

SRI International

AD-A268 682



Final Report • August 3, 1993

Distributed Reasoning and Planning

SRI Project No. ECU - 7363

Prepared by:

**Dr. Kurt G. Konolige, Sr. Computer Scientist
Artificial Intelligence Center**

Prepared for:

**Lcdr. Robert Powell
Scientific Officer
Attention: Code 113D
Office of Naval Research
Ballston Tower One
800 North Quincy Street
Arlington, Va 22217-5660**

This document has been approved
for public release and sale; its
distribution is unlimited.

Approved by:

**Dr. C. Raymond Perrault, Director
Artificial Intelligence Center
Computing and Engineering Sciences Division**

**Dr. Donald D. Nielson, Vice President and Director
Computing and Engineering Sciences Division**

**DTIC
ELECTE
AUG 30 1993
S A D**

93-19572



13107

9 3 8 23 06 5

DISTRIBUTED REASONING AND PLANNING

Final Report for the Period May 1991 - May 1993

July 26, 1993

Kurt Konolige
Karen Myers
Artificial Intelligence Center
Computer Science and Technology Division
SRI International
333 Ravenswood Ave.
Menlo Park, California 94025.

Prepared for:
Office of Naval Research
Information Systems Branch

Accession For	
NTIS CRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By <i>per A257409</i>	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
<i>A-1</i>	

DTIC QUALITY INSPECTED 3

Contents

1	Introduction and Overview	3
1.1	Cooperative Multiagent Domains	4
1.2	The BDI Model	5
1.3	Hybrid Reasoning	5
1.4	Computational Methods for Reasoning about Mental State	7
1.5	Causal Theories	8
2	Hybrid Reasoning	10
2.1	The Hybrid Framework	11
2.1.1	Analogical Subsystem	11
2.1.2	Sentential Subsystem	13
2.2	The Inferential Calculus	14
2.3	Structural Uncertainty	16
2.4	Summary	16
3	Minimal AE Logic	18
3.1	Introduction	18
3.2	Minimal Ideal Introspection	20
3.3	Groundedness, Autoepistemic and Default Logic	21
3.4	Nested Belief	23
3.5	Reflective Reasoning Principles	24
3.6	Conclusion	25
4	Representationalist Theory of Intention	27
4.1	Introduction	27
4.2	Cognitive Structures	29
4.2.1	Rationality Constraints: Intention and Belief	31
4.2.2	Relative Intentions	32
4.3	Conclusion	34

5	A Theory of Causal Reasoning	35
5.1	Causation	35
5.2	Default Causal Nets	37
5.2.1	Causation	37
5.2.2	Definitions and Correlations	39
5.2.3	Normal Conditions	40
5.2.4	Explanations	41
5.2.5	Other Approaches to Explanation	43
5.3	Some Remarks about Causation	44
5.4	Conclusion	46

Chapter 1

Introduction and Overview

This document describes research conducted by SRI International (SRI) on the Office of Naval Research (ONR) project Distributed Reasoning and Planning (Contract N00014-89-C-0095) over the 24-month period from May 1991 to May 1993.

A central focus in artificial intelligence (AI) research is the development of systems that are capable of reasoning about their environments and of planning appropriate courses of action to pursue. Yet for many applications, it is insufficient to rely on a single, autonomous system; instead, a network of physically distributed computer systems must be used. Each of those systems must itself have significant reasoning and planning capabilities, that is, it must be an intelligent agent in its own right. Research in distributed reasoning and planning is concerned with the development of a theoretical framework for designing, building, and managing networks of intelligent agents, which can plan in cooperation with one another to achieve given goals.

Achieving the goal of powerful multiagent systems such as these will require research advances both in the reasoning abilities of the individual agents and in interagent coordination and communication strategies. Most traditional AI planning systems are inadequate for multiagent domains, because of the simplifying assumptions they make. For example, they assume that there is unlimited time available to generate a plan. But multiagent domains are inherently dynamic, because each agent can act independently to change the environment. As a result, the assumption of unlimited reasoning time is invalid, because the environment may change in critical ways during reasoning, and prompt reaction to such change is necessary. Other assumptions made by traditional planning and reasoning systems that are inappropriate for multiagent distributed systems include the assumption of common knowledge of all aspects of the domain as opposed to spatially distributed, specialized knowledge found in distributed sensor domains; centralized construction, distribution, and synchronization of a single joint plan; and unlimited and correct knowledge of other agents' knowledge, intentions, and goals.

In our research, we have therefore been developing improved models of individual planning and reasoning processes that do not rely on these sort of assumptions, as well as techniques that can be used by agents to coordinate their plans. In the last two years, we have concentrated on the following areas:

- Reasoning with analogical representations, particularly with respect to using and learning maps.
- Models of intention and belief that are appropriate in multiagent domains.
- Models of causation. These models are important in deciding what the consequences of agent actions will be.

In the remainder of this section, we first review the characteristics of cooperative multiagent domains. We also review our basic research approach: the adoption of belief-desire-intention (BDI) models of agent state.

In subsequent sections, we provide more detail about the technical aspects of the research we have conducted over the past 24 months. Finally, the appendices contain a selection of the technical papers that we have produced during this period.

1.1 Cooperative Multiagent Domains

Many applications of interest to the Navy require a system of distributed processors, sensors, and effectors. Such systems include distributed command and control, intelligence-gathering networks, squadrons of small submersible vehicles for submarine tracking, and systems for fault isolation and repair of equipment in dangerous or difficult conditions.

Some tasks require sensory surveillance of large areas of land or water by spatially distributed units. Operational considerations often require that these units be able to make local decisions on how best to deploy their sensors. In many cases, it is desirable that the units be mobile, and able to perform other actions. These units will need to communicate with other agents to exchange information about their environments, current goals and intentions. Agents often need to cooperate to process information effectively, and to decide on an effective strategy for obtaining further information.

Multiagent systems have the advantage of tolerance to faults: if one agent cannot effect a given task, then the task may still be achieved with the cooperation of others. These systems also allow for evolutionary development, as single units can be updated independently of the entire system.

Many of these considerations arise even for single agents. In dynamic worlds, an autonomous agent must be able to reason about how it interacts with the changing environment. At any given time, it will have many different goals to accomplish.

Some of these are imposed by fiat as specific tasks that must be accomplished, for example, patrolling a perimeter; others are related to maintaining the agent's own ability to function and gather information about its environment. Often these goals will make conflicting demands on the agent's limited resources, so that the resolution of these conflicts becomes an important part of the reasoning process.

The approach we have taken is to consider a network of autonomous processes cooperating with one another to achieve certain goals. Because the cost of communication is high, it is imperative that each process be capable of reacting intelligently on its own to changes in the environment. This is in contrast to more tightly coupled distributed computational approaches, in which a central scheduler consolidates information and control of the problem-solving process. Furthermore, because of the uncertainty inherent in information gathered from sensory apparatus, and limitations on the functional capabilities of the processes, each process must have a well-developed model of its environment (including the presence of other cooperating agents), and the ability to reason about actions and events in quite complex ways.

1.2 The BDI Model

In designing the agents that inhabit our distributed systems, the approach we have adopted is the belief-desire-intention (BDI) model of mental states [Konolige, 1985c]. This model has been the foundation of our work on *cognitive architectures* [Konolige, 1982; Konolige, 1983; Konolige, 1984; Konolige, 1985a; Konolige, 1985b; Konolige, 1986], including architectures for planning in dynamic environments. It has also been central to our theories of interagent coordination and communication, especially plan recognition [Appelt, 1982; Bratman *et al.*, 1988; Konolige and Pollack, 1989; Pollack, 1986; Helft and Konolige, 1991]. The BDI approach stands in contrast to knowledge-compilation techniques, in which explicit execution-time reasoning is supplanted by compiling into the agent all decisions about what to do in all situations. We believe, along with many other researchers, that such compilation is infeasible for complex environments of the type in which we are interested. However, adopting the BDI approach for dynamic environments commits us to developing efficient reasoning strategies that can function effectively under time constraints.

1.3 Hybrid Reasoning

In our proposal for this project, we stressed the need to develop automatic reasoning systems capable of supporting the complex inferences necessary for reasoning about cognitive state. One of the key aspects of these systems is the ability to do inference based on natural representations of the world. In particular, agents will often want to

refer to and communicate analogical representations of the world. The most familiar form of such representations are maps. Maps contain geometric information about the location of objects such as roads and cities. They also contain symbolic information that can be useful to agents planning a route, for example, the fact that a road is closed in winter. Human agents are very good at combining these diverse forms of information in service of their goals. Typical AI reasoning systems are not so flexible, and have trouble incorporating any analogical representations.

We have initiated a project that addresses the problem of using analogical representations effectively in automated reasoning systems. Analogical representations have the property that their structure embeds properties of the domain being modeled. Maps provide a good example by the manner in which they embed a spatial correspondence with the real world. The class hierarchies used in many knowledge representation systems constitute a nonspatial analogical representation, with the tree structure of the representation mimicking the hierarchical relation of class inclusion. Analogical representations have long been of interest to the AI community, given their dual abilities to encode information in a perspicuous manner and to facilitate efficient manipulations of that information by exploiting embedded structural properties.

During the past two years, we have developed a formal framework for integrating reasoning systems built on analogical and sentential representations; an overview of this work appeared as a paper, "Reasoning with Analogical Representations," in the proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR92). The framework consists of a set of generic operations on analogical structures and deductive rules for applying those operations. The framework supports both reasoning *about* analogical representations, which amounts to a passive extraction of information from analogical structures for use by a sentential reasoning system, and the more general task of reasoning *with* such representations. The latter casts analogical representations in an active role, having them modified as part of the deductive process. The integration rules were proven sound with respect to an introduced model-theoretic semantics for hybrid systems that combine analogical and sentential representations.

To demonstrate the viability of the formal theory, we implemented a prototype hybrid analogical-sentential reasoner. The implementation was built on top of Mark Stickel's KLAUS automated deduction system, using Myers' technology of universal attachments (described by a forthcoming paper "Hybrid Reasoning using Universal Attachment" to appear in the journal *Artificial Intelligence*).

Although our analogical-sentential framework was defined independently of any domain, we have explored its application to the problem of reasoning with maps. Our particular focus has been on the type of maps that our mobile robot can generate from perceptual input as it navigates through an office building. Maps built from sensor information generally have gaps corresponding to areas for which perception

was unable to determine the relevant physical characteristics, due to either faulty sensors, noise or insufficient perceptual cues. Our hybrid analogical-sentential reasoning framework allows a sentential theory describing properties of the environment to be incorporated into the map-making process. Thus, sentences in a logic can be communicated to the robot as a means of improving upon the information provided by perception alone. This communication provides a means of augmenting sensor-based models of the world with information that is beyond the perceptual capabilities of the robot, leading to more accurate and more complete maps.

This work was presented at KR92, as well as several workshops.

1.4 Computational Methods for Reasoning about Mental State

We have continued our work on the development of models of belief and intention, and logics for reasoning about them. There are two separate research lines: ideal belief systems, and the representation of intention. Autoepistemic (AE) logic is a formal system characterizing agents that have complete introspective access to their own beliefs. AE logic relies on a fixed point definition that has two significant parts. The first part is a set of assumptions or hypotheses about the contents of the fixed point. The second part is a set of reflection principles that link sentences with statements about their provability. We have shown, in a paper published in the AAAI conference in July 1992 ("Ideal Introspective Belief") that AE reasoners can be characterized in terms of an assumption set of *negative* beliefs about the world (e.g., "I don't believe that I have an older sister"), together with reflection principles relating beliefs to beliefs about beliefs (e.g., "If I believe X, then I believe that I believe X"). We have shown that AE logic is not an ideal logic, in that negative assumptions are too strong for an ideal introspective agent. This theoretical work can help in analyzing metatheoretic systems in logic programming; this further result was presented in an invited paper at the META92 workshop in Uppsala, Sweden ("An Autoepistemic Analysis of Metalevel Reasoning in Logic Programming").

We are also developing a representationalist logic of intention, which we believe is better suited to the properties of intention than the existing normal modal logic of intentions. Formalizations of cognitive state that include intentions and beliefs have appeared in the recent literature [Cohen and Levesque, 1990; Rao and Georgeff, 1991; Shoham, 1990; Konolige and Pollack, 1989]. With the exception of the work reported here, these have all employed *normal modal logics* (NMLs), that is, logics in which the semantics of the modal operators is defined by accessibility relations over possible worlds. This is not surprising, since NMLs have proven to be a powerful tool for modeling the cognitive attitudes of belief and knowledge. However, we argue that

intention and belief are very different beasts, and that NMLs are ill-suited to a formal theory of intention.

We have developed an alternative model of intention, one that is representationalist, in the sense that its semantic objects provide a more direct representation of cognitive state of the intending agent. We argue that this approach results in a much simpler model of intention than does the use of an NML, and that, moreover, it allows us to capture interesting properties of intention that have not been addressed in previous work. Further, the relation between belief and intention is mediated by the fundamental structure of the semantics, and is independent of any particular choice for temporal operators or theory of action. This gives us a very direct, simple, and semantically motivated theory, and one that can be conjoined with whatever temporal theory is appropriate for a given task.

This work will be presented at IJCAI93.

1.5 Causal Theories

We have also continued our work on proof-theoretic techniques for reasoning about mental state, especially on *abduction*. Simply put, abduction is the process of reasoning from some observation to the best explanation for it. Abduction can be used as a reasoning method for many different kinds of problems. Recently, we have concentrated on its application to causal and default reasoning, important components of reasoning about mental state. In our previous work, we have shown that there are two distinct formalizations for explanatory reasoning. The consistency-based approach treats the task as a deductive one, in which the explanation is deduced from a background theory and a minimal set of abnormalities. The abductive method, on the other hand, treats explanations as sentences that, when added to the background theory, derive the observations. We have shown that there is a close connection between these two formalizations in the context of simple causal theories: domain theories in which a set of sentences are singled out as the explanatorily relevant causes of observations.

In our current work, we expand the idea of abductive inference in causal theories to include defaults. Our theory is unique in that it integrates a formal notion of causality with nonmonotonic reasoning techniques based on default logic and abduction. The main structure of the theory is a default causal net (DCN) representing the causal connections among propositions in the domain. The causal net provides a framework for the two nonmonotonic reasoning techniques of assuming defaults and generating explanations for observations, allowing them to be combined in a principled way. Default causal nets, we claim, offer significant representational advantages over current formal model-based diagnosis theories.

- DCNs distinguish between the strong explanation of the cause of an observation versus the weaker explanation of an excuse for the consistency of the observation.
- Preferences among explanations based on causal relations in DCNs can yield better diagnoses than current model-based theories.
- Because they are based on abductive reasoning, DCNs admit causal influences that are neither normal or abnormal, but neutral.

Some of these advantages accrue because DCNs use an abductive approach to explanation in diagnosis; others, especially the third, are a result of incorporating an explicit causal relation.

This work was presented at the KR92, and accepted for publication in the journal *Annals of Mathematics and Artificial Intelligence*.

Chapter 2

Hybrid Reasoning

This section is based on research by Karen Myers and Kurt Konolige.

Analogical representations have long been of interest to the knowledge representation community [4; 5; 15; 16]. The attraction of analogical representations lies with their ability to store certain types of information that humans can readily process but are problematic for sentential reasoning systems. For this project, we have addressed the problem of using analogical representations effectively in automated deduction systems. The primary outcome of this work is a formal framework for combining analogical and deductive reasoning. The framework consists of a set of generic operations on analogical structures and accompanying inference methods for integrating analogical and sentential information. The capabilities of the framework are demonstrated for the task of reasoning to extend the kind of incomplete maps that might be constructed by a robot operating within an office building. The examples presented here have all been solved automatically by an implementation of the integration framework.

Analogical representations encompass both explicit diagrams (as in [2; 3]) and representation structures that are *diagram-like*. Although this latter class is not easily defined, diagram-like representations share with real diagrams the property of certain structural correspondences with the domain being modeled. It is precisely such correspondences that make analogical representations useful. For example, a two-dimensional street map could be represented by graph-theoretic structures in which nodes correspond to intersections and arcs correspond to road segments. Such a representation is analogical with the world being mapped in two ways. First, paths between nodes in the graph correspond to road connections in the world being modeled. Second, there is a correspondence between the existence of objects in the world and objects in the representation. For example, all roads are represented in the graph; thus, the closure of the set of roads is implicit. In contrast, expressing such closure information sententially would require an explicit statement that the given

roads constitute all roads.

Our work applies equally well to both diagrams and diagram-like structures. For this reason, we will not distinguish further between the two types. The terms *diagram* and *analogical representation* will be used interchangeably in this section.

2.1 The Hybrid Framework

Reasoning with diagrams should not be accomplished by simply translating the diagram contents into a sentential language, nor *vice versa*. Analogical structures provide compact representations of information that is cumbersome to express sententially but generally lack the expressive power of sentential languages. Since sentential theories are a more general representational technology, it is tempting to translate analogical structures into first-order sentences *en masse*. But this strategy would compromise the efficiency of the representation system since the specialized inference mechanisms for the analogical structures are replaced by general-purpose deductive methods; this point is borne out by the experimental results of [9; 11]. Instead, we adopt a hybrid approach in which separate analogical and sentential subsystems coexist and inference rules for translating information between the two are defined.

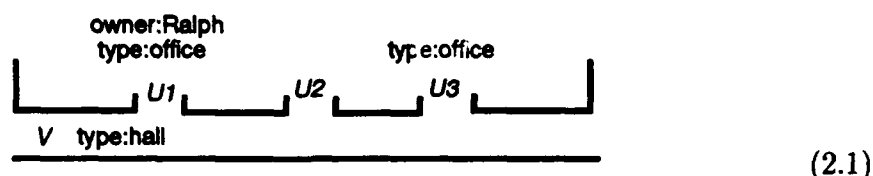
Our hybrid framework is based on a set of generic operations for manipulating analogical structures along with corresponding inference rules that invoke the operations. The operations and rules were chosen for their capacity to increase overall reasoning competency through the appropriate use of analogical information. The framework supports both the incorporation of diagrammatic information into the sentential reasoner and the modification of diagrams to reflect information deduced by sentential reasoning — in other words, both reasoning *about* and *with* diagrams.

2.1.1 Analogical Subsystem

The details of the analogical component will vary for different applications. Our formal framework isolates the integration methods from the specifics of any particular application through the use of an abstract characterization of the information stored in the analogical system.

For example, a typical hallway map used by a mobile robot might contain the

kind of information displayed in the following diagram:



The constants V and U_i are symbolic names assigned to the hallway and the three openings on it in the given scene. These objects and the relationships among them are identified by the robot's perceptual interpretation mechanism, which detects relevant geometric properties and segments sensory input into meaningful units (e.g., groups line segments and intersegment spaces into objects such as corridors and significant openings). We use the term *diagram element* for such objects. Prior knowledge about the scene was used to determine the remainder of the information in this diagram, namely that certain U_i are offices and that the leftmost office belongs to Ralph.

For any particular class of applications, there will be a fixed ontology of elements and a fixed set of properties of interest. We consider two classes of properties: symbolic labels for diagram elements and analogical relations among diagram elements. Formally, we can represent the information about labels and relations for diagram elements that is stored in an analogic representation S as a set of first-order models M_S . While a diagram records only those relationships and elements that are known to exist, each of these *diagram models* constitutes a possible completion of the partial information provided by a diagram. For example, the type of U_2 and the owners of U_2 and U_3 are unspecified in the above diagram; a diagram model would fully specify those relations.

Diagram models consist of a set of analogical relations A and a set of label relations L over a universe U . Each member of A is a binary relation $E_s \times E_s$, with $E_s \subset U$ the set of diagram elements; each member of L is a relation $E_s \times E_l$, with $E_l \subset U$ the set of labels.

For the scene described by (2.1), the diagram elements are $\{V, U_1, U_2, U_3\}$. We choose the label relations $TYPE(u, l)$ and $OWNS(u, l)$, and the analogical relations $BES(u, v)$ (the opening u is next to the opening v) and $INHALL(u, v)$ (opening u is in hall v). The label set contains $\{Closet, Office, Ralph, Paul, Cyril\}$ and possibly other values. The choice of relations and elements is important in determining what information in the analogic structure is abstracted in the hybrid system; here, for example, whether an opening is to the right or left of another opening is apparent from the structure, but not in the models.

A key feature of analogical representations is their capacity to implicitly embody constraints that other representations must make explicit. For example, the map structures embed the following constraints:

- Each opening has at most two adjacent openings.
- Objects can have exactly one type.
- Individuals can own offices but not closets.
- At most one person can own a given office.

These *diagram constraints* can be built into the representation structures directly or into the operations that manipulate the structures, depending on the given implementation. For example, a bit-map representation of (2.1) would embed the first constraint directly through its spatial composition; the third constraint would most likely be enforced by operations that manipulate the structure. Either way, diagram constraints are necessarily reflected in diagram models. For instance, all diagram models for (2.1) can have only one type relation for a given diagram element, because of the second constraint above.

2.1.2 Sentential Subsystem

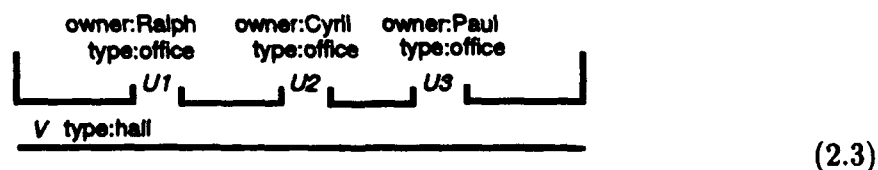
The sentential subsystem employs a first-order language and proof theory. As an example of the expression and use of sentential information relative to diagrams, consider the following statements:

Paul and Cyril have offices in hall V.
Ralph and Paul are not neighbors.

These statements could be translated into formulas for the sentential subsystem such as:

$$\begin{aligned} & RESIDES(Cyril, V) \wedge RESIDES(Paul, V) \\ & \neg NBR(Ralph, Paul) . \end{aligned} \quad (2.2)$$

With respect to diagram (2.1), the first statement implies that U_2 and U_3 are offices, one each owned by Cyril and Paul. This conclusion follows since $\{U_1, U_2, U_3\}$ constitutes the set of all offices in V and *Ralph* is known to own U_1 . Deduction of this result requires information that is implicit in the diagram's structure, namely that each office can be owned by only one individual. With the second statement, the only possible configuration of the scene is:



Our work provides the inferential tools needed to support these kinds of deductions.

In order to determine whether a given integration method behaves in an appropriate fashion, it is necessary to provide a semantic account of the overall hybrid system. Our work provides a model-theoretic account of such criteria. Underlying this work is the introduced notion of the *representational adequacy* of an analogical structure for a sentential theory, which informally indicates that no other analogical structure more accurately represents a given sentential theory. Building on this concept, we present definitions of *soundness*, *derivational completeness* (i.e., completeness with respect to the sentential subsystem) and *diagrammatic completeness* (i.e., completeness with respect to the analogical subsystem) for inference within a hybrid analogical-sentential system.

2.2 The Inferential Calculus

The inferential calculus of the integration framework contains three rules: *reflection*, *evaluation*, and *domain enumeration*. Each rule is defined relative to a domain-independent diagram operation that supports the exchange of information between the sentential and analogical subsystems.

Reflection

The reflection rule sanctions the transfer of information from the sentential to the analogical subsystem. As such, it provides a means of transferring information deduced by the sentential subsystem into the analogical structures. The reflection rule makes use of a collection of *reflection procedures* defined for each analogical predicate. These procedures provide a means of directly inserting information into an analogical structure. Each analogical relationship $P(\bar{x})$ modeled in the analogical structure has an accompanying reflection procedure $\text{INSERT}.P(\bar{x})$ for performing the modifications to the analogical structure.

Evaluation

The evaluation rule sanctions replacement of ground instances of a predicate in formulas of the sentential subsystem by either true or false, in accordance with the information content of the analogical structure. For each analogical relationship $P(\bar{x})$ modeled in the analogical structure, we require an extraction procedure $\text{EVAL}.P(\bar{x})$ for evaluating ground instances relative to a fixed diagram. These *evaluation procedures* provide the sentential reasoner with information about primitive relationships in the analogical structures.

Domain Enumeration

Domain enumeration allows the elimination of quantifiers from formulas in the sentential subsystem in certain cases through the introduction of an appropriate domain of values that covers the relevant instantiations of the quantified variable. This set of values is determined by examination of the current analogical structures.

For example, consider the assertion

$$\exists u. BES(u, U_2) \wedge OWNS(u, Paul) \quad (2.4)$$

relative to diagram (2.1). The interpretation of this formula is that the diagram element owned by Paul is located beside U_2 . The conjunct $BES(u, U_2)$ limits the possibilities for this diagram element: according to (2.1), the element must be either U_1 or U_3 . As such, the formula $OWNS(U_1, Paul) \vee OWNS(U_3, Paul)$ follows from (2.4). Similarly, the universally quantified formula

$$\forall u. INHALL(u, V) \supset TYPE(u, Office) \quad (2.5)$$

can be viewed as a statement about the predicate $TYPE(u, Office)$, with $INHALL(u, V)$ serving as a filter on the set of relevant instantiations of the quantified variable. According to the diagram (2.1), the only values that satisfy $INHALL(u, V)$ are $\{U_1, U_2, U_3\}$ (i.e., the exact closure of $INHALL(u, V)$ is $\{U_1, U_2, U_3\}$). Thus, the conjunction

$$\bigwedge_{d \in \{U_1, U_2, U_3\}} TYPE(d, Office)$$

is equivalent to (2.5) with respect to models for diagram (2.1).

We refer to the technique used above for applying closure information to eliminate quantifiers as *domain enumeration*. Domain enumeration does not apply to all predicate instances containing a quantified variable. The formula $\exists u. \neg BES(u, U_1) \wedge TYPE(u, Closet)$ illustrates this point. In this case, the exact closure for $BES(u, U_1)$ is not an appropriate restriction of the terms of \mathcal{L} ; elimination of the existential quantifier using the exact closure would lead to unsound conclusions.

Our work has identified those cases for which domain enumeration is possible. Application of domain enumeration requires the extraction of *closure information* from the analogical structures. For a predicate $P(\bar{x})$, we use both the set of diagram elements that possibly satisfy $P[x]$ (called *minimal superclosure*) and the set of elements that definitely satisfy $P(\bar{x})$ (the *maximal subclosure*). These closure approximations give minimal upper and maximal lower bounds, respectively, for the precise set of values that satisfy $P(\bar{x})$.

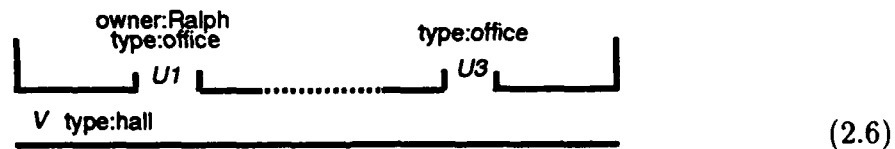
The inference rules of reflection, evaluation and domain enumeration have been proven sound relative to our semantics for analogical-sentential hybrid systems. However, the integration rules are neither derivationally nor diagrammatically complete.

The central problem is that the rules focus on properties of individuals and their relationships with other individuals, failing to account for embedded diagram constraints. We have shown for the propositional case, though, that the refutational version of derivational completeness is achievable using a slight generalization of our methods that draws upon the techniques of theory resolution [18].

2.3 Structural Uncertainty

In diagram (2.1), all objects of relevance (the openings and the hall itself) have been noted, and the analogical relations *BES* and *INHALL* are fully determined. Although there is type and ownership information missing, the *structure* of the diagram is complete. Not all diagrams share this completeness. When generating maps from perceptual input, noise or faulty sensors may both cause objects of interest to go undetected and leave analogical relations only partially determined. In such circumstances, we say that the diagram contains *structural uncertainty*.

The following diagram constitutes a variation on the scene described by (2.1) in which there is structural uncertainty between U_1 and U_3 . Here, both the *BES* and *INHALL* relations are undetermined. Dashed lines indicate regions of structural uncertainty:



Sentential information can also be used to reduce structural uncertainty in diagrams: given the sentences *Ralph and Cyril are neighbors* and *Cyril is Paul's only neighbor*, the diagram (2.3) follows from (2.6). Our integration framework is capable of dealing with structural uncertainty, and can be used to solve the above example.

2.4 Summary

The integration framework has been implemented on top of the KLAUS automated deduction system [19] using the method of *universal attachment* [10; 11] to formulate the integration rules. The system has been successfully applied to problems involving reasoning with maps, including the examples presented here.

While analogical representations have received much attention in recent years from psychologists [6; 7; 8], there have been few advances in understanding the computational aspects of analogical reasoning. Until recently, most computationally oriented work has focused on properties of particular classes of diagrams (e.g., Venn diagrams

[14; 13], Euler circles [17], qualitative reasoning [1; 3; 12], geometry [4]), ignoring more general aspects of reasoning diagrammatically. Our work addresses the broader question of domain-independent inference techniques for reasoning involving analogical representations.

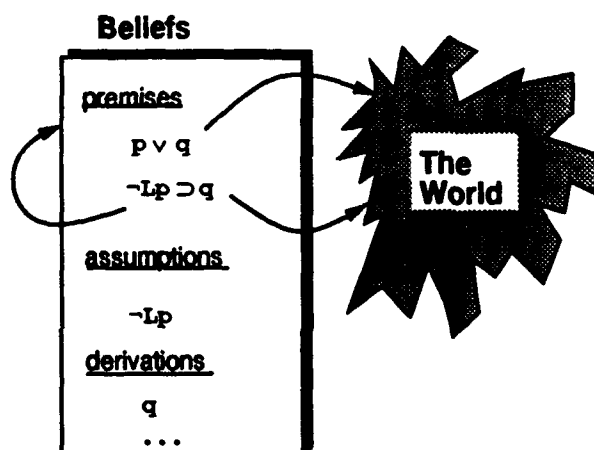
Chapter 3

Minimal AE Logic

This section is based on research by Kurt Konolige.

3.1 Introduction

An important aspect of an agent's reasoning is the ability to introspect about what he believes or does not believe. One of the research lines we pursued was to ask what kind of introspective capability an ideal agent should have. This question is not easily answered, since it depends on what kind of model we take for the agent's representation of his own beliefs. Autoepistemic logic (Moore [Moore, 1985]) uses a sentential or list semantics, which looks like this:



The beliefs of the agent are represented by sentences in a formal language. For simplicity, we consider just a propositional language \mathcal{L}_0 , and a modal extension \mathcal{L}_1 which has modal atoms of the form $L\phi$, where ϕ is a sentence of \mathcal{L}_0 .

The arrow indicates that the intended semantics of the beliefs from \mathcal{L}_0 is given by the real world, for example, the belief q is the agent's judgment that q is true in the real world. Of course an agent's beliefs may be false, so that in fact q may not hold in the world. On the other hand, beliefs of the form $L\phi$ refer to the agent's knowledge of his own beliefs, so the semantics is just the belief set itself.

An agent starts with an initial set of beliefs, the *premises*. Through assumptions and derivations, he accumulates further beliefs, arriving finally at a belief set that is based on the premises. For an agent to be ideally introspective, the belief set Γ must satisfy the following equations:

$$\begin{aligned} &\text{The premises are in } \Gamma. \\ &\phi \in \Gamma \text{ and } \phi \in \mathcal{L}_0 \rightarrow L\phi \in \Gamma \\ &\phi \notin \Gamma \text{ and } \phi \in \mathcal{L}_0 \rightarrow \neg L\phi \in \Gamma \end{aligned} \tag{3.1}$$

Any set Γ from \mathcal{L}_1 that satisfies these conditions, and is closed under tautological consequence, will be called \mathcal{L}_1 -stable (or simply stable) for the premises Γ . The definition and term "stable set" are from Stalnaker [Stalnaker, 1980]. The beliefs are stable in the sense that an agent has perfect knowledge of his own beliefs according to the intended semantics of L , and cannot infer any more atoms of the form $L\phi$ or $\neg L\phi$.

Although an ideal agent's beliefs will be a stable set containing his beliefs, not just any such set will do. For example, if the premises are $\{p \vee q\}$, one stable set is $\{p \vee q, p, Lp, L(p \vee q), \dots\}$. This set contains the belief p , which is unwarranted by the premises. The constraint of making the belief set stable guarantees that the beliefs will be introspectively complete, but it does not constrain them to be soundly based on the premises. Moore recognized this situation in formulated autoepistemic logic; his solution was to ground the belief set by making every element derivable from the premises and some assumptions about beliefs. The reason he needed a set of assumptions is that negative introspective atoms (of the form $\neg L\phi$) are not soundly derivable from the premises alone. For example, consider the premise set $\{\neg Lp \supset q, p \vee q\}$. We would like to conclude $\neg Lp$, since there is no reasonable way of coming to believe p . But an inference rule that would allow us to conclude $\neg Lp$ would have to take into account all possible derivations, including the results of its own conclusion. This type of circular reasoning can be dealt with by adding a set of assumptions about what we expect *not* to believe, and checking at the end of all derivations that these assumptions are still valid.

In autoepistemic logic, a belief set T is called *grounded in premises* A if all of its members are tautological consequences of $A \cup LT_0 \cup \neg L\bar{T}_0$, where $LT_0 = \{L\phi \mid \phi \in T \cap \mathcal{L}_0\}$, and $\neg L\bar{T}_0 = \{\neg L\phi \mid \phi \in \mathcal{L}_0 \text{ and } \phi \notin T\}$. This concept of groundedness is fairly weak, since it relies not only on assumptions about what isn't believed ($\neg L\bar{T}_0$), but also about what is (LT_0). In this paper we consider belief sets that use only

assumptions $\neg L\bar{T}_0$ in forming the belief set T . Everything else in the belief set will follow deductively (and monotonically) from the premises A and the assumptions $\neg L\bar{T}_0$. In some sense $\neg L\bar{T}_0$ is the minimal set of assumptions that we can use in this manner; for every smaller set, we have to resort to nonmonotonic rules, such as negation-as-failure [Lloyd, 1987], in order to form a stable set. For this reason we call a belief set grounded in A and $\neg L\bar{T}_0$ *ideally grounded*.

Ideally grounded logics are similar to the modal nonmonotonic logics defined in [McDermott, 1982; Shvarts, 1990; Marek *et al.*, 1991], but allow an agent to make fewer assumptions about his own beliefs. The main difference is that ideally grounded logics are more grounded in the premises than modal nonmonotonic logics, and in general will have fewer unmotivated extensions.

In the rest of this chapter we explore ideally grounded belief sets from the perspective of introspective reflection principles. We are able to characterize the minimal set of principles that will yield a stable set of beliefs, and also (once nested belief operators are introduced) the maximal ones. The resultant family of introspective logics fills in a hierarchy between strongly and moderately grounded autoepistemic logic [Konolige, 1988], and suggests that the moderately grounded fixed-point is the best system for an ideal agent with perfect awareness of his beliefs.

3.2 Minimal Ideal Introspection

In this and the following section we restrict the language to \mathcal{L}_1 , containing no nesting of the belief operator. This presents a simple system to explore the consequences of ideal introspection.

An ideally grounded introspective agent determines his belief set using the following fixed-point equation:

$$T = \{\phi \mid A \cup \neg L\bar{T}_0 \vdash_S \phi\}, \quad (3.2)$$

where S is some system of inference rules. Any set T that satisfies this equation will be called an *ideally grounded extension* of A . The set $T_0 = T \cap \mathcal{L}_0$ is the *kernel* of T .

In the remainder of this section we consider the minimal set of rules S that guarantees a stable belief set for T . Because a stable set is closed under tautological consequence, the rules S must contain a complete set of propositional rules. In addition, whenever ϕ is in the belief set, we want to infer $L\phi$. The following two rules fulfill these conditions.

Rule Taut. From the finite set of sentences X infer ϕ , if ϕ is a tautological consequence of X .

Rule Reflective Up. From ϕ infer $L\phi$, if $\phi \in \mathcal{L}_0$.

PROPOSITION 3.2.1 *Let RN be the rules Taut and Reflective Up. Every RN-extension of A is a \mathcal{L}_0 stable set containing A .*

PROPOSITION 3.2.2 *If for every set $A \subseteq \mathcal{L}_1$, the S -extension of A is an \mathcal{L}_1 stable set containing A , then Taut and Reflective Up are admissible rules of S .*

These two propositions show that the rules RN form the minimal logic for ideally grounded agents, in the sense that RN extensions produce stable belief sets, and they must be included in any system that produces such sets. Further, every RN extension of A is *minimal for A* : there is no stable set S containing A such that $S_0 \subset T_0$.

PROPOSITION 3.2.3 *Every RN extension of A is a minimal stable set for A .*

Thus, we have shown that two simple rules, Taut and Reflective Up, are sufficient to guarantee an ideal introspective agent.

3.3 Groundedness, Autoepistemic and Default Logic

In this section we relate ideally grounded extensions to their close relatives, default logic and AE extensions. Ideal groundedness is somewhat weaker than default logic and strongly grounded AE extensions, but stronger than moderately grounded ones.

Simple as it is, the system RN is almost equivalent to default logic [Reiter, 1980]. It is not quite as strongly grounded as the latter; while there exists a translation from DL to RN that preserves extensions, the inverse translation fails in a few cases.

We will assume that the reader is familiar with DL. A default theory (W, D) consists of a set of first-order sentences W and a set of defaults D of the form

$$\alpha : \beta_1, \dots, \beta_n / \gamma.$$

Here only the propositional case will be considered, but extending the results to first-order languages is straightforward (as long as no quantifying-in is allowed, e.g., sentences of the form $Qx.L\phi(x)$).

To get a translation to RN, simply take W and add a translation of each default rule, as follows:

$$A = W \cup \{L(\alpha \wedge \alpha) \wedge \neg L\neg\beta_1 \dots \supset \gamma \mid \alpha : \beta_1, \dots / \gamma \in D\}. \quad (3.3)$$

Note the form of the first modal atom: $L(\alpha \wedge \alpha)$, rather than $L\alpha$. Since the beliefs of an agent are closed under tautological consequence, this amounts to the same constraint on beliefs; however, the difference is important for finding extensions, as will be made clear shortly.

PROPOSITION 3.3.1 *U is the kernel of an RN extension of A iff it is a DL extension of $\langle W, D \rangle$.*

This is a simple translation of DL into a minimal AE logic. It is the same as the translation in [Konolige, 1988] (except for the use of $\alpha \wedge \alpha$ instead of α), but there it was necessary to limit the extensions of the AE logic to strongly grounded ones, a syntactic method based on the form of the premises. No such method is needed here.

To get autoepistemic logic, we need to include more assumptions about beliefs in the fixed point equation 3.2. Let us define *open RN extensions* as solutions of the equation

$$T = \{\phi \mid A \cup LT_0 \cup \neg L\bar{T}_0 \vdash_{RN} \phi\}, \quad (3.4)$$

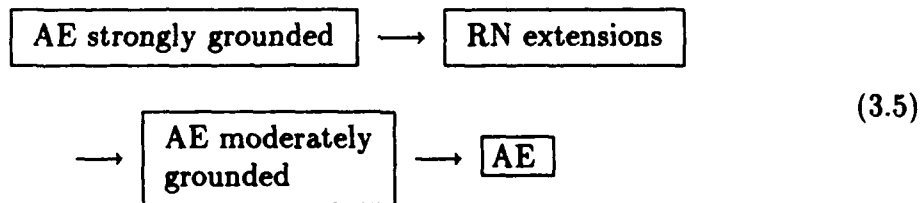
where LT_0 is the set $\{L\phi \mid \phi \in T_0\}$. Actually, the presence of the Up rule is redundant here. From results in [Konolige, 1988], it is easy to show the following proposition.

PROPOSITION 3.3.2 *T is an open RN extension of A iff it is the kernel of an AE extension of A .*

The kernel of an AE extension is just the part of the extension from \mathcal{L}_0 . The kernel completely determines the extension.

So the basic difference between AE and default logic is based on the groundedness of the extensions, that is, AE logic lets an agent assume belief in a proposition α , and use that assumption to derive the very same proposition as part of the final set of beliefs. In default logic, all derivations must be ideally grounded, so that assumptions are of the form $\neg L\phi$.

The circular reasoning possible in AE logic was noted in [Konolige, 1988], and two increasingly stronger notions, moderate and strong groundedness, were defined as a means of throwing out extensions that exhibit such reasoning. RN extensions are related to these systems by the following diagram:



The arrows indicate inclusion of the logics: AE logic admits the most extensions, and AE strongly grounded the fewest.

3.4 Nested Belief

So far we have preferred to forego the complications of beliefs about beliefs, using the language \mathcal{L}_1 that contains no nesting of modal operators. This language and its semantics can be extended in a straightforward way. Let \mathcal{L} be the propositional modal language formed from \mathcal{L}_0 by the recursive addition of atoms of the form $L\mu$, with $\mu \in \mathcal{L}$.

The semantic equations for a stable set (3.1) are modified to take away the restriction of beliefs being in \mathcal{L}_0 :

$$\begin{aligned} &\text{The premises are in } \Gamma. \\ &\phi \in \Gamma \rightarrow L\phi \in \Gamma \\ &\phi \notin \Gamma \rightarrow \neg L\phi \in \Gamma \end{aligned} \tag{3.6}$$

Any set from \mathcal{L} that satisfies these conditions, and is closed under tautological consequence, will be called a stable set for A (in contrast to \mathcal{L}_1 -stable, which does not consider nested modal atoms).

Consider a premise set A that is drawn from \mathcal{L}_1 , as before. In every RN extension of A there is complete knowledge of what facts are believed or disbelieved, i.e., $L\phi$ or $\neg L\phi$ is present for every nonmodal ϕ . The addition of the nested modal atoms should make no difference to this picture, except to reflect the presence of the belief atoms in the correct way. So, for example, if La is in an RN extension S , then LLa should be in the extension when we consider \mathcal{L} ; and similarly $L\neg La$ should be present if $\neg La$ is not in S . This much is easily accomplished by removing the restriction on Reflective Up, and giving it its usual name from modal logic.

Rule Necessitation. From ϕ infer $L\phi$.

This rule will add positive modal atoms; but we need also to add negative ones. For example, if La is in an extension, and the extension is consistent, then $\neg La$ is not in it, and this fact should be reflected in the presence of $\neg L\neg La$. In fact we want to infer $\neg L\mu$ for *every* sentence μ that will not be in the extension, given that we have full knowledge of the belief atoms from \mathcal{L}_1 . Suppose that there is a sentence $La \vee \neg Lb \vee c$ that is not in S , where c is a nonmodal sentence. This implies that, for stable S , $\neg La \in S$, $Lb \in S$, and $\neg Lc \in S$. So from these latter sentences we should infer $\neg L(La \vee \neg Lb \vee c)$. This is what the following rule does.

Rule Fill. From $L\alpha_i$, $\neg L\beta_j$, $\neg L\gamma$, and $\mu \supset (\bigvee_i L\alpha_i \vee \bigvee_j \neg L\beta_j \vee \gamma)$, infer $\neg L\mu$.

The system NRN consists of the rules Taut, Necessitation, and Fill. The basic properties of NRN extensions are that they are minimal stable sets, the rules are essential, and they are conservative extensions of RN fixed points.

PROPOSITION 3.4.1 *If for every set $A \subseteq \mathcal{L}$, the S -extension of A is a stable set containing A , then *Taut*, *Necessitation*, and *Fill* are admissible rules of S .*

PROPOSITION 3.4.2 *Every NRN extension of A is a stable set for A .*

Extensions that are stable sets are also minimal, as for the nonnested language.

PROPOSITION 3.4.3 *If the rules S are sound with respect to stable sets, and the S -extension of A is a stable set, then it is a minimal stable set for A .*

The nested extensions are conservative with respect to nonnested ones.

PROPOSITION 3.4.4 *If $A \subseteq \mathcal{L}_1$, then the kernel of every RN extension is the kernel of an NRN extension, and conversely, the kernel of every NRN extension is the kernel of an RN extension.*

Finally, the *Fill* rule is redundant in the presence of the *K* axiom schema.

PROPOSITION 3.4.5 *The rule *Fill* is admissible in any system containing *K*, *Taut* and *Necessitation*.*

Because nested modal atoms are propositionally distinct from nonnested ones, it is possible to derive new translations from default logic to sentences of \mathcal{L} such that all extensions are strongly grounded and hence equivalent to default logic extensions. There are many ways to do this; all that is required is to translate from $\alpha : \beta/\gamma$ to a sentence in which α and β are put under different nestings of modal operators that correspond to the single nesting semantics. For example, three such translations are:

$$\begin{aligned} a) & \quad LL\alpha \wedge \neg L\neg\beta \supset \gamma \\ b) & \quad L\alpha \wedge \neg LL\neg\beta \supset \gamma \\ c) & \quad L\alpha \wedge L\neg L\neg\beta \supset \gamma \end{aligned} \tag{3.7}$$

3.5 Reflective Reasoning Principles

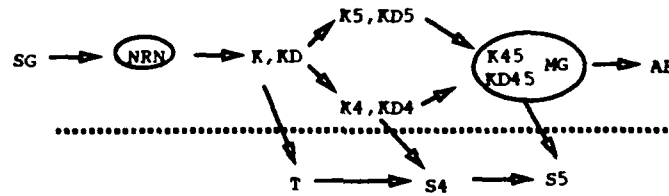
The systems RN and NRN are minimal rules that might be used by an agent reasoning about its own beliefs. They have the nice characteristic of giving minimal stable sets, and so are somewhere between strongly and moderately grounded. But are there other reflective reasoning principles that could be incorporated? In this section we will give a partial answer to this question by examining several standard modal axiomatic schemata, and showing how some of them are appropriate as general reasoning principles, while others must be regarded as specific assumptions about the relation of beliefs to the world.

The most well-known modal schemata are the following.

$$\begin{aligned}
 K. & L(\phi \supset \psi) \supset (L\phi \supset L\psi) \\
 T. & L\phi \supset \phi \\
 D. & L\phi \supset \neg L\neg\phi \\
 4. & L\phi \supset LL\phi \\
 5. & \neg L\phi \supset L\neg L\phi
 \end{aligned}
 \tag{3.8}$$

Different modal systems can be constructed by combining the different modal schemata with the inference rules Taut and Necessitation. Using our previous definition of inclusion, we show the following relations among the different versions of *S*-extensions.

PROPOSITION 3.5.1 *The following diagram gives all the inclusion relations of ideally grounded extensions based on the modal systems formed from the schemas *K*, *T*, *D*, 4, and 5.*



The top half are systems whose extensions are all subsets of AE logic. SG stands for strongly grounded AE extensions, and MG for moderately grounded. The minimal ideally grounded system is NRN, and the maximum is K45 or KD45, which is equivalent to MG (see [Konolige, 1988]). An ideal introspective agent would use KD45 extensions, which we call ideal extensions. Note that the schema *D* does not make any difference as far as ideally grounded extensions are concerned; in effect, the agent cannot use reasoning about self-belief to detect an incoherence in his beliefs.

3.6 Conclusion

We have presented the minimal logic (NRN) that an ideal introspective agent should use. It is minimal in the sense that the agent makes a minimal set of assumptions about his own beliefs, and employs a minimal set of rules necessary to guarantee that his beliefs are stable. An ideal introspective reasoner may enjoy more powerful rules of introspection, for example the modal schemas 4 and 5, but he should keep the assumptions about his beliefs to a minimum. The schema *T* is not a sound axiom

for an introspective agent, but can be used to characterize a contingent connection between beliefs and the world.

The concept of ideally grounded extensions first appeared in [Konolige, 1988], where the system KD45 was presented and proven equivalent to moderately grounded AE extensions.¹ Fixpoints of the systems T, S4 and S5 were introduced under the name of nonmonotonic ground logics in [Tiomkin and Kaminski, 1990], and it was shown that the S5 logic was nondegenerate and consistent, i.e., does not reduce to monotonic S5, and always has an extension.

Ideally grounded logic might be employed in an analysis of metatheoretic systems, such as the DEMO and SOLVE predicates in logic programming [Bowen and Kowalski, 1982; Costantini, 1990]. Using a predicate to represent provability can cause problems with syntax and consistency (see [des Rivières and Levesque, 1986] for some comments). Instead, this research suggests using a modal operator, and defining a theory by the fixed point definition (3.2). Some appropriate notion of negation-as-failure would be used to generate the assumptions, and the rest of the fixed point could be calculated using the reflection rules.

¹ A slightly different fixed-point was used because of a technical difference in the form of monotonic inference in modal systems.

Chapter 4

Representationalist Theory of Intention

This section is based on research by Kurt Konolige and Martha Pollack.

4.1 Introduction

Formalizations of cognitive state that include intentions and beliefs have appeared in the recent literature [Cohen and Levesque, 1990; Rao and Georgeff, 1991; Shoham, 1990; Konolige and Pollack, 1989]. With the exception of the work presented here, these have all employed *normal modal logics* (NMLs), that is, logics in which the semantics of the modal operators is defined by accessibility relations over possible worlds. This is not surprising, since NMLs have proven to be a powerful tool for modeling the cognitive attitudes of belief and knowledge. However, we argue that intention and belief are very different beasts, and that NMLs are ill-suited to a formal theory of intention.

We therefore present an alternative model of intention, one that is representationalist, in the sense that its semantic objects provide a more direct representation of cognitive state of the intending agent. We argue that this approach results in a much simpler model of intention than does the use of an NML, and that, moreover, it allows us to capture interesting properties of intention that have not been addressed in previous work. Further, the relation between belief and intention is mediated by the fundamental structure of the semantics, and is independent of any particular choice for temporal operators or theory of action. This gives us a very direct, simple, and semantically motivated theory, and one that can be conjoined with whatever temporal theory is appropriate for a given task.

NMLs have been widely and successfully used in the formalization of belief. It

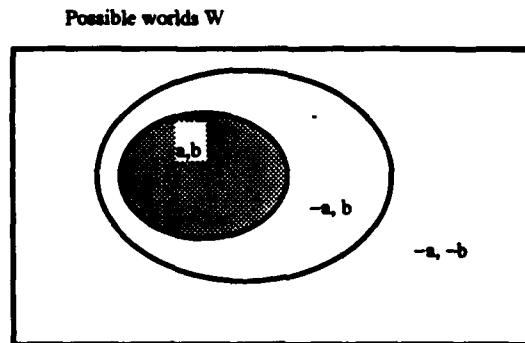


Figure 4.1: A Venn diagram of two scenarios.

is largely as a result of this success that researchers have adopted them in building models of intention. However, we argue that these logics are inappropriate to models of intention:

- The semantic rule for normal modal operators is the wrong interpretation for intention. This rule leads to the confusion of an intention to do ϕ with an intention to do any logical consequence of ϕ , called the *side-effect problem* [Bratman, 1987]. A simple and intuitively justifiable change in the semantic rule makes intention side-effect free (and nonnormal).
- Normal modal logics do not provide a means of relating intentions to one another. Relations among intentions are necessary to describe the means-end connection between intentions.

These problems do not mean we have to abandon possible worlds. In fact, with the right semantics, possible worlds are an intuitively satisfying way of representing future possibility and intention for an agent. We note that intentions divide the possible futures into those that the agent wants or prefers, and those he does not. Consider the diagram of Figure 4.1. The rectangle represents the set of possible worlds W . The *scenario* for a proposition a is the set of worlds in W that make a true: the shaded area in the diagram. An agent that has a as an intention will be content if the actual world is any one of those in the shaded area, and will be unhappy if it is any unshaded one. The division between wanted and unwanted worlds is the important concept behind scenarios. For example, consider another proposition b that is implied by a (for concreteness, take a = "I get my tooth filled," and b = "I feel pain.") If we just look at interpretations within the shaded area, a and b both hold, and so cannot be distinguished. But the complement of these two propositions is different. A world in the area $\neg a, b$, in which the agent feels pain but does not have his tooth pulled, is an acceptable world for the intention b , but not for a . So the interpretation rule for

intention must take into account the complement of the intended worlds. As we will see in Section 4.2, this makes intention a nonnormal modal operator. It also makes it side-effect, abstraction, and conjunction free, whether we choose realism or weak realism.

The representationalist part of the model comes in representing the mental state of the agent using scenarios. *Cognitive structures*, containing elements representing intentions and the relationship among intentions, are used for this purpose.

4.2 Cognitive Structures

Our model of intention will have two components: possible worlds that represent possible future courses of events, and *cognitive structures*, a representation of the mental state components of an agent. We introduce complications of the model in successive sections. To begin, we define the simplest model, a static representation of primary or “top-level” intentions. Primary intentions do not depend on any other intentions that the agent currently has.

The concept of intention is intimately connected with choosing among courses of future action. In the model, courses of action are represented by possible worlds. Each possible world is a complete history, specifying states of the world at all instants of time. We assume there is a distinguished moment *now* in all worlds that is the evaluation point for statements.

To talk about contingent and necessary facts, we use the modal operators \Box and \Diamond . The possibility operator \Diamond expresses the existence of a world with a given property. $\Diamond\phi$ says that there is a world (among W) for which ϕ is true. \Diamond is used to specify the background of physically possible worlds under which reasoning about intention takes place, and will be important in describing the structure of a given domain. The necessity operator $\Box\phi$ is defined as $\neg\Diamond\neg\phi$.

A key definition is the concept of scenario.

DEFINITION 4.2.1 *Let W be a set of possible worlds, and ϕ any sentence of \mathcal{L} . A scenario for ϕ is the set*

$$M_\phi = \{w \in W \mid w, W \models \phi\}.$$

A scenario for ϕ identifies ϕ with the subset of W that make ϕ true.

A cognitive structure consists of the background set of worlds, and the beliefs and intentions of an agent.

DEFINITION 4.2.2 *A cognitive structure is a tuple $\langle W, \Sigma, \mathcal{I} \rangle$ consisting of a set of possible worlds W , a subset of W (Σ , the beliefs of the agent) and a set of scenarios over W (\mathcal{I} , the intentions of the agent).*

We extend the language by adding the modal operators B for belief and I for intentions. The beliefs of an agent are taken to be the sentences true in all worlds of Σ . For simplicity, we often write Σ as a set of sentences of \mathcal{L}_\Box , so that M_Σ is the corresponding possible worlds set.

The beliefs of an agent are always possible, that is, they are a subset of the possible worlds. This also means that an agent cannot be wrong about necessary truths. A more complicated theory would distinguish an agent's beliefs about what is possible from what is actually possible. The key concept is that intentions are represented with respect to a background of beliefs about possible courses of events (represented by \Diamond), as well as beliefs about contingent facts (represented by B). The following theorems are key facts about belief:

$$\begin{aligned} B(\phi) &\supset \Diamond\phi \\ B(\Box\phi) &\equiv \Box\phi \end{aligned} \tag{4.1}$$

Of course, beliefs about contingent facts can still be false, since the real world does not have to be among the believed ones. The B operator represents all futures the agent believes might occur, including those in which he performs various actions or those in which he does nothing. The beliefs form a background of all the possibilities among which the agent can choose by acting in particular ways.

The third component of a cognitive structure for an agent, an intention structure, is a set of scenarios M_ϕ . Intuitively, an agent's intention structure will include one scenario for each of his primary intentions. We write \mathcal{I} as a set of sentences of \mathcal{L}_\Box , where each sentence ϕ stands for its scenario M_ϕ . $I(\phi)$ is true just in case ϕ is equivalent to some proposition $\psi \in \mathcal{I}$, given the background structure W .

PROPOSITION 4.2.1 *For any structure $\langle W, \Sigma, \mathcal{I} \rangle$,*

$$\langle W, \Sigma, \mathcal{I} \rangle \models I(\phi) \quad \text{iff} \quad \exists \psi \in \mathcal{I}. W \models \Box(\phi \equiv \psi).$$

The I operator is true precisely of the individual top-level intentions the agent has. It is not subject to closure under logical consequence or under the agent's beliefs. To see this, consider the cognitive structure $\langle W, \Sigma, \{a\} \rangle$, i.e., the agent has the single intention to perform a . Assume that a logically implies b , but not the converse, i.e.,

$$W \models \Box(a \supset b) \wedge \Diamond(b \wedge \neg a).$$

Then $M_a \neq M_b$, because there is a world in which b is true but a is not. From the semantics of I , we have

$$\langle W, \Sigma, \{a\} \rangle \models I(a) \wedge \neg I(b)$$

This shows that I is not closed with respect to valid consequence.

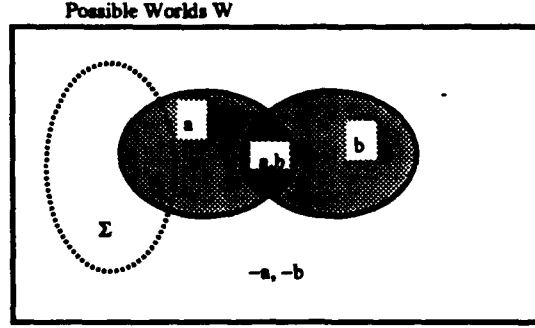


Figure 4.2: A Venn diagram of belief and intention.

4.2.1 Rationality Constraints: Intention and Belief

So far we have not related the agent's intentions to his beliefs. Consider the diagram of Figure 4.2, for which the cognitive structure is $\langle W, \Sigma, \{a, b\} \rangle$. The agent's two intentions are jointly possible, since the overlapping area contains at least one world in which they both hold. However, based on the contingent facts of the situation, the agent does not believe that they will actually occur, since his beliefs, given by the set Σ , fall outside the overlap area. A rational agent will not form intentions that he does not believe can be jointly executed. Further, intentions should be nontrivial, in the sense that the agent intending ϕ should not believe that ϕ will occur without the intervening action of the agent. To enforce rationality, we define the following conditions on cognitive structures.

DEFINITION 4.2.3 A cognitive structure $\langle W, \Sigma, \mathcal{I} \rangle$ is admissible iff it is achievable:

$$\exists w \in \Sigma. \forall \phi \in \mathcal{I}. w \in M_\phi$$

and nontrivial:

$$\forall \phi \in \mathcal{I}. \exists w \in \Sigma. w \notin M_\phi.$$

This condition leads immediately to the following consequences.

PROPOSITION 4.2.2 *These sentences are valid in all admissible structures.*

$\neg I(a \wedge \neg a)$	<i>Consistency</i>
$I(a) \wedge I(b) \supset \Diamond(a \wedge b)$	<i>Joint Consistency</i>
$I^*(a) \supset \Diamond a$	
$I^*(a) \supset B\Diamond a$	<i>Realism</i>
$I(a) \supset \neg B(\neg a)$	<i>Epistemic Consistency</i>
$I(a) \wedge I(b) \supset \neg B(\neg(a \wedge b))$	<i>Joint Epistemic Consistency</i>
$I^*(a) \supset \neg B\neg(a)$	
$I(a) \supset \neg B(a) \wedge \neg B(\neg a)$	<i>Epistemic Indeterminacy</i>

A rational agent, characterized by achievable structures, does not believe that his joint intentions represent an impossible situation: this is the theorem of Joint Epistemic Consistency. This theorem can be stated using either reading of intention.

In addition, the nontriviality condition on models means that the agent does not believe that any one of his intentions will take place without his efforts (Epistemic Indeterminacy). Recall that the B operator represents all futures the agent believes might occur, including those in which he performs various actions or does nothing. The beliefs form a background of all the possibilities among which the agent can choose by acting in particular ways. If in all these worlds a fact ϕ obtains, it does no good for an agent to form an intention to achieve ϕ , even if it is an action of the agent, because it will occur without any choice on the part of the agent. So, for example, if the agent believes he will be forced to act at some future point, perhaps involuntarily (e.g., by sneezing), it is not rational for the agent to form an intention to do that.

4.2.2 Relative Intentions

One of the primary characteristics of intentions is that they are structures: agents often form intentions relative to pre-existing intentions. That is, they "elaborate" their existing plans. A plan can be elaborated in various ways. For instance, a plan that includes an action that is not directly executable can be elaborated by specifying a particular way of carrying out that action; a plan that includes a set of actions can be elaborated by imposing a temporal order on the members of the set; and a plan that includes an action involving objects whose identities are so far underspecified can be elaborated by fixing the identities of one or more of the objects. As Bratman [Bratman, 1987, p.29] notes, "[p]lans concerning ends embed plans concerning means and preliminary steps; and more general intentions ... embed more specific ones." The distinction between these two kinds of embedding recurs in the AI literature. For instance, Kautz [Kautz, 1990] identifies two relations: (1) *decomposition*, which relates a plan to another plan that constitutes a way of carrying it out (means and

preliminary steps), and (2) *abstraction*, which relates a specific plan to a more general one that subsumes it. It is useful to have a term to refer to the inverse relation to abstraction: we shall speak of this as *specialization*.

Both kinds of elaboration are represented in the cognitive structure by a graph among intentions. The graph represents the means-ends structure of agent intentions. For example, suppose the agent intends to do a by doing b and c . Then the cognitive structure contains the graph fragment $M_b, M_c \rightarrow M_a$. As usual, in the cognitive structure we let the propositions stand for their associated scenarios.

DEFINITION 4.2.4 *An elaborated cognitive structure consists of a cognitive structure and an embedding graph \rightarrow among intentions: $\langle W, \Sigma, \mathcal{I}, \rightarrow \rangle$. The graph is acyclic and rooted in the primary intentions.*

Remarks. The reason we need both primary intentions and the graph structure is that, while every root of the graph must be a primary intention, primary intentions can also serve as subordinate intentions. Consider the masochistic agent with a tooth cavity: he both intends to feel pain, and intends to get his tooth filled. His cognitive structure would be:

$$\{W, \{a \supset b\}, \{a, b\}, a \rightarrow b\}.$$

Also note that a scenario of the graph may serve to elaborate more than one intention; Pollack [Pollack, 1991] calls this overloading.

The embedding graph \rightarrow is the most strongly representationalist feature of the model. It represents the structure of intentions in a direct way, by means of a relation among the relevant scenarios. A normal modal logic is incapable of this, because its accessibility relation goes from a single world (rather than a scenario) to a set of possible worlds.

As with primary intentions, we can specify suitable rationality constraints for subsidiary intentions. The key constraint has to do with the means-end relation. An agent should believe that if the elaboration is achieved, the original intention will be also. Consider the diagram of Figure 4.3, in which the agent has the intention to achieve a by achieving b ; for concreteness, take the example of calling the telephone operator by dialing 0. There can be possible worlds in which b does not lead to a : for example, in using the internal phone system of a company. The correct rationality condition for an agent is that he believe, in the particular situation at hand, that achieving b will achieve a . This is represented by the set Σ of belief worlds, in which $b \supset a$ holds. We call a model *embedded* if it satisfies this constraint on belief and intention structure.

While the embedding graph semantics is simple, it leads to interesting interactions in the statics of intention and belief. For example, in plan recognition it can be used to determine if a recognized plan is well-formed. It is also critical to the theory of

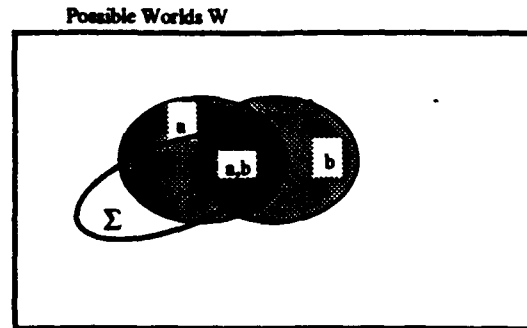


Figure 4.3: Means-ends intentions and belief.

the dynamics of intention and belief. We have a preliminary theory of this dynamics expressed as a default system.

4.3 Conclusion

We have concentrated on the static relation between intention and belief, and shown how the relationship between these two can be represented simply by an appropriate semantics. The static formalism is useful in a task such as plan recognition, in which one agent must determine the mental state of another by using partial information.

More complex applications demand a *dynamic* theory, which is really a theory of belief and intention revision. The formalism of cognitive structures can be extended readily to time-varying mental states, by adding a state index to the model. However, the theory of revision is likely to be complicated, even more so than current belief revision models [Gärdenfors and Makinson, 1990], and will probably involve elements of default reasoning.

Chapter 5

A Theory of Causal Reasoning

5.1 Causation

Knowledge of causation is an important part of commonsense reasoning. We use cause-and-effect analysis to understand everything from why we caught the flu to how to make a video recorder save our favorite TV show. If causation is so ubiquitous in reasoning about and affecting everyday events, it might also be useful to employ this concept in a formal theory of diagnosis. Surprisingly, the best-known such theory, *model-based diagnosis* [Reiter, 1987], does not. We argue in this paper that importing a formal notion of causation into model-based diagnosis leads to a better theory, solving some significant representational and inference problems.

What benefits can an explicit encoding of causation bring to diagnostic theories? There are at least three possible areas:

- Problem structuring
- Explanations
- Computation

The first, problem structure, is the most important, and underlies the other two. It is clear that in everyday reasoning we use the concept of cause and effect to structure our interpretations of the observations we make, to understand how events occur and how we can affect them. This representational issue is the main focus of the paper, and just below we present an example motivating our viewpoint.

The second item, explanations, is important whenever a diagnostic system must communicate its results to an end user. In answering questions about how a conclusion was reached, it is not acceptable for a system to state:

X is 13 and Y was 12 and the system equation predicts that Z will be 18.

This kind of "explanation" will not be helpful: it does not give a user insight into the domain in terms that he is familiar with, i.e., causal relations.

Finally, there are computational issues. By giving a structure to the domain, one that usually has a strong acyclic bias, causal relations can focus the computational task. Some examples of the benefits that can result are in the theory of Bayes nets [Pearl, 1988] and in using causal approximations to physical theories [Nayak, 1992a; Nayak, 1992b]. Although we give some computational methods at the end of this paper, these are mostly to touch base with previous work in model-based diagnosis, and we have not yet explored the computational ramifications of the theory.

A causal default theory should address two tasks: prediction and explanation. Prediction is the process of deriving the course of events from initial conditions. Prediction is useful in many ways, for example, in planning one's actions. What happens if I don't pay my telephone bill on time? Knowing the consequences of this action can help decide whether to perform it or not. Another way prediction is used is to set up expectations in testing. An electronics engineer may apply an input to a circuit, expecting it to generate a certain output if it is working correctly.

The second task is explanation: from observed effects, infer what could have caused that effect. Typical here are applications such as plan recognition and diagnosis of complex systems. In plan recognition, one tries to infer the intentions of someone through observation of her actions: *Why did the train conductor ask if I had a passport?* Understanding the relation of actions to intentions is important in any cooperative task, and especially in communication [Cohen *et al.*, 1990]. Diagnosis is a similar kind of task, except that one is trying to figure out possible explanations for a system not behaving as expected: *Why does the copier always jam when I put in transparency paper?* Finding the answer to this question can help in fixing the problem.

We have developed a theory that integrates causal and default reasoning within a first-order framework. Both the normal function of a system, and full or partial information about its fault modes can be represented. The main structure of the theory is a default causal net (DCN) representing the causal connections among propositions in the domain. Default causal nets, we claim, offer significant representational advantages over current formal model-based diagnosis theories.

- DCNs distinguish between the strong explanation of the cause of an observation versus the weaker explanation of an excuse for the consistency of the observation.
- Partial fault models are allowed; information about fault modes can lead to stronger explanations, but complete information is not required.
- Preferences among explanations based on causal relations in DCNs can yield

better diagnoses than current model-based theories.

- Because it is based on abductive reasoning, DCNs admit causal influences that are neither normal or abnormal, but neutral.

Some of these advantages accrue because DCNs use an abductive approach to explanation in diagnosis; others, especially the third, are a result of incorporating an explicit causal relation.

5.2 Default Causal Nets

Default causal nets (DCNs) are a formal structure that encode the concepts of causation, correlation, and defaults. They consist of a causal theory R , a definitional theory D , and a correlation or integrity theory I . In addition there are distinguished sets of propositions C (the primitive causes) and N (the normal conditions). The term "net" is used in analogy with Bayesian nets, because the main structuring concept is the causal relation embodied in R .

DEFINITION 5.2.1 (DEFAULT CAUSAL NET)

A default causal net is a tuple $\langle R, D, I, C, N \rangle$, where R is a Horn theory, D and I are first-order theories, and C and N are disjoint sets of atoms.

5.2.1 Causation

Formally, we understand causation to be a primitive relation among propositions. By "primitive" we mean that, as far as DCNs are concerned, the causation relation is part of the parameterization of the net, and is not derived from any other concepts. This is unlike the approach of Shoham [Shoham, 1987], for example, in which a theory of causation is developed by reducing it to other concepts. Our approach leaves unanswered questions about how to identify causation in a given domain, the relation of causation to time, and various other difficulties about the nature and properties of causation.

To represent the causal relation, we use a definite clause theory R over a first-order language \mathcal{L} . This theory consists of a set of implications

$$a_1 \cdots a_n \supset b.$$

where each of a_i and b is a ground atom of \mathcal{L} . If A is a set of propositions, then we say that an atom b is caused by A if there is a proof of b from A in R ; we write this as $A \vdash_R b$. A is a minimal cause for b if there is no other cause A' for b such that $A' \subset A$.

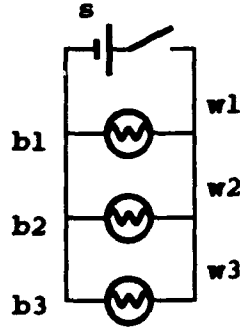


Figure 5.1: Three bulbs with a switch

EXAMPLE 5.2.1 A variation of the 3-bulb example is diagrammed in Figure 5.1. There is a switch that can be either *open* or *closed*. For each of the other components c_i , the proposition $ok(c_i)$ means that the component is working, and $ab(c_i)$ that it is broken. The theory R is:

$$\begin{aligned}
 &closed, ok(s), ok(w_1), ok(b_1) \supset on(b_1) \\
 &closed, ok(s), ok(w_1), ok(w_2), ok(b_2) \supset on(b_2) \\
 &closed, ok(s), ok(w_1), ok(w_2), ok(w_3), ok(b_3) \supset on(b_3) \\
 &open \supset off(b_1) \\
 &open \supset off(b_2)
 \end{aligned} \tag{5.1}$$

We have not listed any fault models, although we could. Here is a partial fault model that we will use in some examples.

$$\begin{aligned}
 &ab(b_1) \supset off(b_1) & ab(b_2) \supset off(b_2) \\
 &ab(w_1) \supset off(b_1) & ab(w_2) \supset off(b_2)
 \end{aligned} \tag{5.2}$$

The partial fault model is also part of the relation R , since it represents causation in the abnormal functioning of the device. The primitive causes C are $\{open, closed, ab(c_i)\}$.

Note that there can be causes other than the normal or abnormal functioning of a component. This is useful in representing neutral situations, e.g., the switch is not normally either closed or open, but can be hypothesized as either in order to explain the observations. The propositions $ok(x)$ are not listed as primitive causes; they are normal conditions, explained below.

The important part of the causal relation is that it captures the functional dependence of the domain variables. If we want to turn b_1 on, then we can close the switch

and make sure that s , w_1 , and b_1 are working correctly. On the other hand, we cannot make b_1 be on as a means of causing the switch to close. Of course, if we observe b_1 to be on, then we can infer that the switch is closed; but it is not possible to *plan* to change the position of the switch by the primitive action of making the bulb be on. This illustrates the difference between a causal relation and a merely correlational one. Unlike material implication, the causal relation is asymmetric and does not contrapose: given that c causes d , it is not necessarily the case that $\neg d$ causes $\neg c$. Deduction in a definite clause theory is one way to represent the asymmetric causal relation.

5.2.2 Definitions and Correlations

Besides causation, there are other types of relations connecting propositions. Definitional information relates propositions that have defined relations within a domain, e.g., "a 40-watt bulb is a type of bulb" or "abnormal is the opposite of normal." Definitions can obviously interact with causation, since from "a broken 40-watt bulb caused the problem" we can infer "a broken bulb caused the problem." For our purposes, we limit definitions to information about complementary propositions. Definitional relations are represented by a first-order theory D ; for the bulbs example of Figure 5.1, it contains the propositions:

$$\begin{aligned} open &\equiv \neg closed \\ ab(c_i) &\equiv \neg ok(c_i) \\ on(b_i) &\equiv \neg off(b_i) \end{aligned} \tag{5.3}$$

If $p \vdash_D \neg q$, then we say that q is the complement of p , and write it as \bar{p} .

Information about co-occurrences is another form of non-causal information in a domain, e.g., "Whenever I clean my car it rains." Correlations can be used to make predictions, but do not contribute to causal explanations. Correlations are represented by a first-order theory I (for *integrity* theory). All causation and definition relations are also correlational. We enforce this restriction by demanding that $R \subseteq I$ and $D \subseteq I$.

EXAMPLE 5.2.2 Continuing the bulbs example, suppose we know that whenever b_1 is off and is not broken, the other bulbs must be off, too. We represent this as

$$off(b_1) \wedge ok(b_1) \supset off(b_2) \wedge off(b_3) \in I \tag{5.4}$$

Correlations may come from many different sources. As in the case of this example, there may be underlying but unknown causes that link several propositions. Or we

may have experiential knowledge that is the converse of causation: whenever the road is wet, it normally rained the previous night.

A proposition q is correlationally inferred from a set of propositions A if it follows logically from the correlational theory and A ; we write $A \vdash_I q$. For example, $off(b_2)$ is inferred from $A = \{ok(b_1), off(b_1)\}$ in the above example, but it is not caused by A . If A causes q , then it also infers q , since $R \subseteq I$. Note that, unlike the case with the causal relation, the material conditional can be used for "backwards" inference, e.g., if $on(b_2)$ is true, we can infer that one of $ab(b_1)$ or $on(b_1)$ is true by using the contrapositive of Equation 5.4.

5.2.3 Normal Conditions

Normal conditions are propositions that are normally assumed to hold. They generally represent either the normal functioning of a component, or a complex set of conditions, e.g., "if the key is turned and *everything is normal*, the car will start." Formally, normal conditions are a set of ground atoms N that are not primitive causes. Primitive causes are hypotheses that incur a cost to assume; normal conditions are "free" and assumed to hold by default.

EXAMPLE 5.2.3 Continuing the bulbs example, we let the set of normal conditions $N = \{ok(c_i)\}$. In this case, the normal conditions just describe the correct functioning of the components. We can define other types of normal conditions, for example to relate causation among abnormal components. Suppose that normally when b_1 is on, it causes b_2 to fail. We would write:

$$n \wedge on(b_1) \supset ab(b_2) \quad (5.5)$$

as part of the causal theory R , where $n \in N$ is a new proposition reflecting a normal causal relation between b_1 and b_2 . As we will show later, such causal relations can be used to specify priorities among explanations.

Identifying normal conditions is the key to default reasoning in causal theories. We seek to explain a set of observations by hypothesizing causes that are as "normal" as possible, that is, conflict with the fewest normal conditions.

It is helpful to view the causal relation and normal conditions as a directed graph. For example, the normal functioning of the bulbs with the switch closed (Equation 5.1) and the failure mode just given (Equation 5.5) can be diagrammed as in Figure 5.2. The arrows show the causal connections among propositions, annotated with their normal conditions (for simplicity we have omitted some irrelevant normal conditions). The circled arrow indicates that bulb 1 being on is the cause of an abnormal condition with bulb 2. The causal directionality is clear from the diagram.

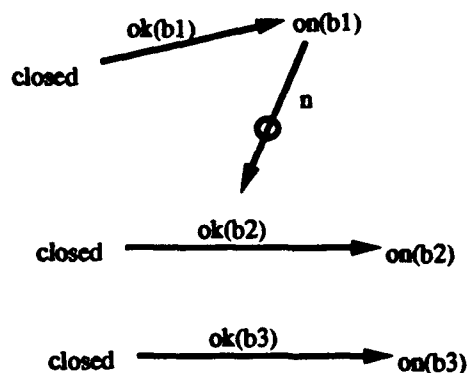


Figure 5.2: Causal directionality

The choice of what conditions are assumed to be “normal” or part of the causal background is an important part of the information provided by the application developer. Depending on the task and the level of expertise of the developer, very different choices could be made, even in the same domain. For example, a typical driver might infer that turning the key causes the car to start, given the normal condition that the car is ok. A car mechanic might have a more detailed causal view: turning the key and having a charged battery causes the car to start, assuming the starter motor is working correctly.

5.2.4 Explanations

We now have all of the elements necessary to develop the inference operation of explanation within DCNs.

DEFINITION 5.2.2 (EXPLANATION)

An explanation for an observation set O is a set of causes and normal conditions $A \subseteq C \cup N$ such that $A \vdash_R O$ and $A \cup O \not\vdash_I \perp$.

EXAMPLE 5.2.4 To illustrate the concept of explanation, we consider the bulbs theory containing the normal causal rules (5.1) together with the fault model (5.2). The fault model is necessary to provide interesting explanations of nonnormal behavior. Suppose we make the observation that bulb b_1 is not lit: $off(b_1)$. There are several explanations for this proposition.

$open, ok(s), ok(b_1), ok(w_1)$
 $closed, ok(s), ab(b_1)$
 etc.

There are usually many explanations for a given observation set, and we seek intuitively preferred explanations. To find these, we filter all explanations by a two-step process.

1. Normal explanation: those explanations that satisfy a maximal set of normal conditions.
2. Ideal explanation: normal explanations that have a minimal number of primitive causes.

The concept of a normal explanation is complicated by the presence of causation. An abnormal condition may be caused by the explanation; when this happens, we say that the normal condition is *exempted*. A normal explanation should consistently either include or exempt as many normal conditions as possible. Here we are using the concept of causation to structure the defaults. If a normal condition is not contained in an explanation, it counts against the explanation, *unless* the corresponding abnormal condition is exempted.

DEFINITION 5.2.3 (ADJUNCT)

Let A be an explanation for observation set O . The adjunct of A is a set of normal conditions defined as follows.

- *If the complement \bar{x} of a normal condition x is in A , then x is in the adjunct.*
- *If a normal condition x is not in A , and $A \not\models_R \bar{x}$, then x is in the adjunct.*

A normal explanation for O is one whose adjunct does not strictly contain the adjunct of any other explanation for O . An ideal explanation is a normal one that is subset-minimal in the primitive causes.

EXAMPLE 5.2.5 As in the previous example, consider the bulbs theory (5.1) together with the fault model (5.2). Again, if we make the observation that bulb b_1 is off, we have several candidates for normal explanations:

Explanation	Adjunct
$ok(s), open, ok(w_1), ok(b_1) \dots$	none
$ok(s), ab(w_1), ok(b_1) \dots$	$ok(w_1)$
$ok(s), ok(w_1), ab(b_1) \dots$	$ok(b_1)$

Of these, the minimal adjunct is the first. This is the normal and ideal explanation of $off(b_1)$: the switch is open, and all components are normal.

This example illustrates one property of normal explanations: as many normal conditions are assumed to hold as possible. The switch can be either open or closed; if we assume that it is open, then we have an explanation for b_1 being off that is consistent with the normal functioning of the circuit. Any other explanation will force us to assume that some component is functioning abnormally. So, normal explanations consist of a set of primitive causes that explain the observations, and at the same time respect our ideas about what normally occurs as much as possible.

In this example, there were no interesting causal relations between normal conditions. In the definition of adjunct, we used the principle of *causal exemption*: if an abnormal condition is caused by the hypothesized explanation, then it is exempted from consideration in finding the "most normal" explanation. The following example illustrates this point.

EXAMPLE 5.2.6 Consider the same fault model as in Example 5.2.5 with an initial condition *closed* and the additional causal rule (5.5): $n \wedge on(b_1) \supset ab(b_2)$. There are several candidates for normal explanations of $\{off(b_2)\}$:

Explanation	Adjunct
$n, ok(s), ok(w_1), ok(b_1), ok(w_2) \dots$	none
$n, ok(s), ok(w_1), ok(b_1), ab(w_2) \dots$	$ok(w_2)$
$ok(s), ok(w_1), ok(b_1), ok(w_2), ab(b_2) \dots$	$ok(b_2), n$
etc.	

Of these, the first is the only normal explanation, and hence ideal. The reason it has an empty adjunct is that the normal conditions and *closed* cause $on(b_1)$, which in turn causes $ab(b_2)$, exempting the normal condition $ok(b_2)$. Every other explanation violates at least one normal condition without exempting it. This makes intuitive sense: if the switch is closed, we expect b_1 to be on, causing b_2 to be broken and off.

This example illustrates how directionality in the causal relation is important in producing causal preferences among explanations. Referring back to Figure (5.2), it is easy to see from following the causal arrows that *closed*, $ok(b_1)$ and n are a cause of $ab(b_2)$. On the other hand, *closed* and $ok(b_2)$ are inconsistent with n and $ok(b_1)$, but they do not cause the complement of either of these normal conditions.

5.2.5 Other Approaches to Explanation

Although we have concentrated on the application of DCNs to diagnosis, they provide a general framework for representing causation and explanation. Causation can

be used as a unifying concept to understand various perspectives on diagnosis: excusing vs. explaining observations, correlation vs. causation, and the integration of normal conditions with explanatory causes. Although many of these issues have been dealt with separately in the literature, there have been few attempts to draw them together into a single framework, and the issues are often obscured by the formal or computational paradigm. There are many formal nonmonotonic systems that provide similar capabilities, although they are not phrased in terms of causation, e.g., Poole's THEORIST [Poole, 1988]. DCNs are distinguished by providing a coherent account of causation, correlation, and default conditions. Perhaps the closest system is Geffner's theory of causal and conditional reasoning [Geffner, 1989], which also takes causation as a primitive concept, and ties together explanation, defaults, and causation. He provides a complex but plausible formal account of these concepts, using a modal expression $C\alpha$ to represent " α is caused." Although the formalisms differ, there are many points of similarity between this work and his. Perhaps the major difference is that the roots of DCNs are default logic and abductive inference, and thus there are natural computational methods using the ATMS.

A good test of the DCN framework is the application to reasoning about events. We have started this task, and it appears that the problems of causation, explanation, and prediction in an event calculus can be treated within the DCN framework. The approach is similar to that of Shanahan [Shanahan, 1989], but the formal machinery is more general, and includes causation.

5.3 Some Remarks about Causation

Perhaps the weakest point of the DCN approach is that the theory of causation is not well developed. Since causation is treated as a proof-theoretic concept, there are some obvious problems (or, one might say opportunities) that arise. We discuss some of these here; a more detailed treatment can be found in [Konolige, 1991].

First, there is a deliberate sloppiness about stating propositions in the causal relation. Most of the ones used in this paper are statements about particular properties, e.g., the switch is closed or the light is on. But causation also involves events: "closing the switch caused the light to go on." We are trying to be as noncommittal as possible about the ontology of events and propositions, whether states of the world can be allowed as causes, how to specify the time of events, and so on. Any consistent defensible set of choices will do.

The second point is that a definite clause of R must specify *all* and *only* the propositions governing an effect. Closing the switch turns on the bulbs only if they are ok and the wires are intact. Of course, in any real-world situation there will be an inordinate number of such conditions, so any default causal theory will be relative

to a set of background assumptions that do not enter into the theory. The choice of these assumptions is conventional.

It is important that only the relevant propositions participate in the causal relation. If we add an irrelevant proposition to the antecedent of a clause, the relation would still be useful in the sense that conjunction of the antecedents produces the desired effect, but it would be misleading in implying that all the antecedents were necessary. In producing explanations, minimal causal antecedents are required in the causal relation to ensure that explanations do not contain irrelevant propositions.

The role of primitive causes is to define the propositions over which, in some sense, we can exercise direct control. The point at which we choose to define primitive causes is partly a matter of convention. Often bodily movements are taken to be the ultimate primitive causes, but this viewpoint is unnecessarily restrictive. Any well-defined event or condition that we can reliably bring about will suffice for a primitive cause, as long as the purpose of producing explanations is to give a set of conditions that account for the observed facts, and over which we have control.

One way to understand the causation relation R is as a provability relation. The provability relation is composed from individual inference steps combined into a tree; in the same way, the causation relation is specified by combining definite clause inference steps into a proof. Like classical provability, causation is monotonic:

$$\text{If } A \vdash_R c \text{ and } B \supset A, \text{ then } B \vdash_R c$$

and cumulative:

$$\text{If } A \vdash_R c \text{ and } B, c \vdash_R d, \text{ then } A, B \vdash_R d.$$

As we have stated, the important part of the causal relation is that it captures the functional dependence of the domain variables; this is the main difference between a causal relation and a merely correlational one. The asymmetry of causation is represented by the asymmetry of inference in a definite clause theory.

These remarks leave open the question of whether, in a particular instance, it is possible to have a causation relation that is symmetric for two propositions, or more generally to have one that is cyclic, containing a loop that leads from a proposition back to the same proposition. Other commitments may answer this question: for instance, assuming that causes always precede their effects in time forces the causal relation to be acyclic. The definite clause theory itself does not enforce any acyclic condition.

There are some further complications in defining a causal relation that we will mention here, without offering any definitive solutions. The first is that of inferred causation. We mentioned this briefly in proposing the definitional theory in Section 5.2.2. We use only a simple form of definitions to represent complements; any full-fledged theory of causation should at least take into account abstraction relations among propositions, e.g., "A 40-watt bulb is a type of bulb."

Another problem arises when our knowledge of the causation relation is partial. We have already remarked that we may know only a subset of the actual causation relation. Other kinds of uncertainty also exist. For example, suppose we know that dialing the number "911" connects one with either the police or the fire department, but we don't know which. The action of dialing 911 is completely determinate, it's just that we don't know the exact outcome. To express epistemic uncertainty of this kind, it is necessary to describe the causation relation in an appropriate language. If we let c stand for the action of dialing 911, d for calling the police, and e for calling the fire department, then our knowledge is expressed by the statement:

Either $c \vdash_R d$ or $c \vdash_R e$.

DCNs are not expressive enough to state this; a language that talks about causation, such as Geffner's [Geffner, 1989], would be necessary.

5.4 Conclusion

We have developed a theory of causation in the presence of defaults about normally occurring conditions. The theory is based on structure called Default Causal Nets, which integrate causal, correlational, and definitional information. These nets can be used to generate predictions and explain observations.

We have argued that preferences among explanations can be based on noting how causation and defaults interact. Such preferences seem to follow commonsense reasoning based on causal knowledge. In model-based diagnosis, any assumptions about causation and defaults are implicit in the representation of components as being normal or abnormal, and the search for diagnoses is based on abnormal components. Such a view, we argue, is representationally restrictive, and does not give a deep enough analysis about how defaults interact. For example, although we can state relations among abnormalities in the domain, these relations do not necessarily lead to intuitively correct preferences among diagnoses in the consistency-based approach, because material implications within the framework are not treated as causal relations.

Bibliography

- [Appelt, 1982] Douglas E. Appelt. Planning natural-language utterances to satisfy multiple goals. Technical Report 259, SRI International, 1982. Also appears as a Stanford University Ph.D. thesis.
- [Bowen and Kowalski, 1982] K. A. Bowen and R. A. Kowalski. *Amalgamating Language and Metalanguage*, pages 153–172. Academic Press, 1982.
- [Bratman *et al.*, 1988] Michael E. Bratman, David J. Israel, and Martha E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4(4), 1988. Also will appear in J. Pollock and R. Cummins, eds., *Philosophy and AI: Essays at the Interface*, MIT Press, Cambridge, MA.
- [Bratman, 1987] Michael E. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.
- [Cohen and Levesque, 1990] Philip R. Cohen and Hector Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990.
- [Cohen *et al.*, 1990] P. R. Cohen, J. Morgan, and M. E. Pollack, editors. *Intentions in Communication*, Cambridge, MA, 1990. MIT Press.
- [Costantini, 1990] S. Costantini. Semantics of a metalogic programming language. *International Journal of Foundations of Computer Science*, 1(3), 1990.
- [des Rivières and Levesque, 1986] J. des Rivières and H. Levesque. The consistency of syntactical treatments of knowledge. In Joseph Y. Halpern, editor, *Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 115–130. Morgan Kaufmann, 1986.
- [Gärdenfors and Makinson, 1990] P. Gärdenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In M. Vardi, editor, *Conference on Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann, 1990.

- [Geffner, 1989] Hector Geffner. *Default Reasoning: Causal and Conditional Theories*. PhD thesis, Department of Computer Science, University of California at Los Angeles, 1989.
- [Helft and Konolige, 1991] N. Helft and K Konolige. A theory of plan recognition based on abduction and relevance. Working paper, 1991.
- [Kautz, 1990] Henry A. Kautz. A circumscriptive theory of plan recognition. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, MA, 1990.
- [Konolige and Pollack, 1989] Kurt Konolige and Martha Pollack. Ascribing plans to agents: Preliminary report. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989.
- [Konolige, 1982] Kurt Konolige. A first order formalization of knowledge and action for a multiagent planning system. In J. E. Hayes, D. Michie, and Y-H. Pao, editors, *Machine Intelligence 10*. Ellis Horwood Limited, Chichester, England, 1982.
- [Konolige, 1983] Kurt Konolige. A deductive model of belief. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, 1983. Universität Karlsruhe.
- [Konolige, 1984] Kurt Konolige. *A Deduction Model of Belief and its Logics*. PhD thesis, Stanford University, 1984.
- [Konolige, 1985a] Kurt Konolige. Belief and incompleteness. In Jerry R. Hobbs and Robert C. Moore, editors, *Formal Theories of the Commonsense World*, pages 359-404. Ablex Publishing Corporation, Norwood, New Jersey, 1985.
- [Konolige, 1985b] Kurt Konolige. A computational theory of belief introspection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Los Angeles, 1985.
- [Konolige, 1985c] Kurt Konolige. Experimental robot psychology. Technical Note 363, SRI Artificial Intelligence Center, Menlo Park, California, 1985.
- [Konolige, 1986] Kurt Konolige. *A Deduction Model of Belief*. Pitman Research Notes in Artificial Intelligence, 1986.
- [Konolige, 1988] Kurt Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35(3):343-382, 1988.

- [Konolige, 1991] Kurt Konolige. What's happening: elements of commonsense causation. In *Proceedings of the International Conference on Cognitive Science*, San Sebastian, Spain, May 1991.
- [Lloyd, 1987] J. W. Lloyd. *Foundations of Logic Programming*. Springer-Verlag, Berlin, 1987.
- [Marek et al., 1991] W. Marek, G. F. Schwarz, and M. Truszczyński. Modal non-monotonic logics: ranges, characterization, computation. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, Cambridge, MA, 1991.
- [McDermott, 1982] Drew McDermott. Non-monotonic logic II. *Journal of the ACM*, 29:33-57, 1982.
- [Moore, 1985] Robert C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25(1), 1985.
- [Nayak, 1992a] P. Pandurang Nayak. Causal approximation. In *Proceedings of the Conference of the American Association of Artificial Intelligence*, pages 703-709, Menlo Park, CA, 1992. AAAI Press/MIT Press.
- [Nayak, 1992b] P. Pandurang Nayak. Order of magnitude reasoning using logarithms. In *Proceedings of the International Conference on Knowledge Representation and Reasoning*, pages 201-210, San Mateo, CA, 1992. Morgan Kaufmann.
- [Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
- [Pollack, 1986] Martha E. Pollack. Inferring domain plans in question-answering. Technical Report 403, SRI International, Menlo Park, CA, 1986. Also appears as a University of Pennsylvania PhD thesis.
- [Pollack, 1991] Martha E. Pollack. Overloading intentions for efficient practical reasoning. *Nous*, 1991. To appear.
- [Poole, 1988] David Poole. A methodology for using a default and abductive reasoning system. Technical report, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1988.
- [Rao and Georgeff, 1991] Anand S. Rao and Michael P. Georgeff. Modelling rational agents within a bdi-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, Cambridge, MA, 1991.

- [Reiter, 1980] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81-132, 1980.
- [Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57-96, 1987.
- [Shanahan, 1989] Murray Shanahan. Prediction is deduction but explanation is abduction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989.
- [Shoham, 1987] Yoav Shoham. *Reasoning about Change*. MIT Press, Cambridge, MA, 1987.
- [Shoham, 1990] Yoav Shoham. Agent-oriented programming. Technical Report STAN-CS-90-1335, Stanford University, Palo Alto, CA, 1990.
- [Shvarts, 1990] G. F. Shvarts. Autoepistemic modal logics. In *Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA, 1990.
- [Stalnaker, 1980] R. C. Stalnaker. A note on nonmonotonic modal logic. Department of Philosophy, Cornell University, 1980.
- [Tiomkin and Kaminski, 1990] M. Tiomkin and M. Kaminski. Nonmonotonic default modal logics. In *Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA, 1990.
- [1] B. V. Funt. Problem-solving with diagrammatic representations. *Artificial Intelligence*, 13, 1980.
- [2] G. W. Furnas. Formal models for imaginal deduction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 662-669. Lawrence Erlbaum, 1990.
- [3] F. Gardin and B. Meltzer. Analogical representations of naive physics. *Artificial Intelligence*, 38:139-159, 1989.
- [4] H. Gelernter. Realization of a geometry-theorem proving machine. In E. A. Feigenbaum and J. Feldman, editors, *Computers and Thought*. McGraw-Hill, New York, 1963.
- [5] P. J. Hayes. Some problems and non-problems in representation theory. In *Proceedings of the AISB Summer Conference*, pages 63-79, University of Sussex, 1974.
- [6] P. N. Johnson-Laird. Mental models in cognitive science. *Cognitive Science*, 4:71-115, 1980.

- [7] S. M. Kosslyn. *Image and Mind*. Harvard University Press, Cambridge, MA, 1980.
- [8] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11:65-99, 1987.
- [9] K. L. Myers. *Universal Attachment: An Integration Method for Logic Hybrids*. PhD thesis, Stanford University, 1991.
- [10] K. L. Myers. Universal attachment: An integration method for logic hybrids. In J. A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*. Morgan Kaufmann, 1991.
- [11] K. L. Myers. Hybrid reasoning using universal attachment. *Artificial Intelligence*, 1992. To appear.
- [12] E. P. Novak, Jr. and W. C. Bulko. Understanding natural language with diagrams. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 465-470, 1990.
- [13] S.-J. Shin. An information-theoretic analysis of valid reasoning with Venn diagrams. In J. Barwise, J. M. Gawron, G. Plotkin, and S. Tutiya, editors, *Situation Theory and its Applications*, volume 2. 1991.
- [14] S.-J. Shin. *Valid Reasoning and Visual Representation*. PhD thesis, Dept. of Philosophy, Stanford University, 1991.
- [15] A. Sloman. Interactions between philosophy and AI. *Artificial Intelligence*, 2, 1971.
- [16] A. Sloman. Afterthoughts on analogical representation. In *Proceedings of Theoretical Issues in Natural Language Processing*, 1975.
- [17] K. Stenning and J. Oberlander. Spatial containment and set membership. In J. Barnden and K. Holyoak, editors, *Analogical Connections*. 1992.
- [18] M. E. Stickel. Automated deduction by theory resolution. *Journal of Automated Reasoning*, 1(4), 1985.
- [19] M. E. Stickel. The KLAUS automated deduction system. In *Proceedings of the Ninth International Conference on Automated Deduction*, 1988.

Reasoning with Analogical Representations

Karen L. Myers Kurt Konolige
Artificial Intelligence Center
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
myers@ai.sri.com konolige@ai.sri.com

Abstract

Analogical representations have long been of interest to the knowledge representation community. Such representations provide compact encodings of information that can be cumbersome to represent and inefficient to manipulate in sentential languages. In this document, we address the problem of using analogical representations effectively in automated deduction systems. The primary contribution is a formal framework for combining analogical and deductive reasoning. The framework consists of a set of generic operations on analogical structures and accompanying inference methods for integrating analogical and sentential information. The capabilities of the framework are demonstrated for the task of reasoning to extend incomplete maps. The examples presented here have all been solved automatically by an implementation of the integration framework.

1 Introduction

Analogical representations have long been of interest to the knowledge representation community [8, 9, 22, 23]. The attraction of analogical representations lies with their ability to store certain types of information that humans can readily process but are problematic for sentential reasoning systems. Although the power of analogical representations has been acknowledged for many years, little progress has been made in understanding how to exploit the computational advantages that these representations can provide.

Analogical representations encompass both explicit diagrams (as in [6, 7]) and representation structures that are *diagram-like*. Although this latter class is not easily defined, diagram-like representations share with real diagrams the property of certain structural correspondences with the domain being modeled. It is pre-

cisely such correspondences that make analogical representations useful. For example, a two-dimensional street map could be represented by graph-theoretic structures in which nodes correspond to intersections and arcs corresponds to road segments. Such a representation is analogical with the world being mapped in two ways. First, paths between nodes in the graph corresponds to road connections in the world being modeled. Second, there is a correspondence between the existence of objects in the world and objects in the representation. For example, all roads are represented in the graph; thus, the closure of the set of roads is implicit. In contrast, expressing such closure information sententially would require an explicit statement that the given roads constitute all roads.

The work described in this paper applies equally well to both diagrams and diagram-like structures. For this reason, we will not distinguish further between the two types. The terms *diagram* and *analogical representation* will be used interchangeably throughout the document.

While analogical representations have received much attention in recent years from psychologists [10, 11, 12], there have been few advances in understanding the computational aspects of analogical reasoning. Until recently, most computationally-oriented work has focused on properties of particular classes of diagrams (e.g., Venn diagrams [21, 20], Euler circles [24], qualitative reasoning [5, 7, 18], geometry [8]), ignoring more general aspects of reasoning diagrammatically. This document addresses the broader question of domain-independent inference techniques for reasoning involving analogical representations. The work encompasses both reasoning *about* and *with* diagrams. The former involves extraction of information from a diagram and amounts to a passive use of diagrams; the latter further supports modifications to diagrams as a result of the reasoning process, thus constituting an active use of diagrams.

Reasoning with and about diagrams should not be accomplished by simply translating the diagram contents

into a sentential language, nor *vice versa*. Analogical structures provide compact representations of information that is cumbersome to express sententially but generally lack the expressive power of sentential languages. Since sentential theories are a more general representational technology, it is tempting to translate analogical structures into first-order sentences *en masse*. But this strategy would compromise the efficiency of the representation system since the specialized inference mechanisms for the analogical structures are replaced by general-purpose deductive methods; this point is borne out by the experimental results of [13, 15]. Here, we adopt a hybrid approach in which separate analogical and sentential subsystems co-exist and inference rules for translating information between the two are defined.

Our hybrid framework is based on a set of generic operations for manipulating analogical structures along with corresponding inference rules that invoke the operations. The operations and rules were chosen for their capacity to increase overall reasoning competency through the appropriate use of analogical information. The framework supports both the incorporation of diagrammatic information into the sentential reasoner and the modification of diagrams to reflect information deduced by sentential reasoning; in other words, both reasoning about and with diagrams.

One particular class of analogical representations to which we apply our work is that of office-building maps. We are currently using an implementation of our framework in the construction of a hybrid map-learning architecture for the SRI mobile robot [16]. For concreteness, we focus on examples from this application; the work, however, applies to all types of analogical structures. Our examples employ schematic map diagrams whose exact representations are left unspecified; the choice of a particular representation is immaterial to the research presented here.

2 The Hybrid Framework

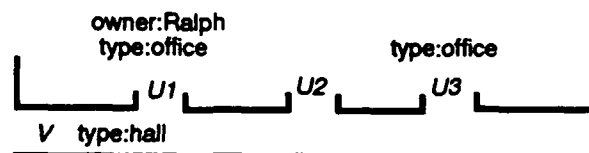
In this section, we describe the analogical and sentential subsystems along with criteria for their integration. Specific integration rules are presented in Sections 3-4.

2.1 Analogical Subsystem

The details of the analogical component will vary for different applications. Our formal framework isolates the integration methods from the specifics of any particular application through the use of an abstract characterization of the information stored in the analogical system. Since we are interested in reasoning with maps, we employ examples from that domain here.

A typical hallway map used by a mobile robot might

contain the kind of information displayed in the following diagram:



(1)

The constants V and U_i are symbolic names assigned to the hallway and the three openings on it in the given scene. These objects and the relationships among them are identified by the robot's perceptual interpretation mechanism, which detects relevant geometric properties and segments sensory input into meaningful units (*e.g.*, groups line segments and intersegment spaces into objects such as corridors and significant openings). We use the term *diagram element* for such objects. Prior knowledge about the scene was used to determine the remainder of the information in this diagram, namely that certain U_i are offices and that the leftmost office belongs to Ralph.

For any particular class of applications, there will be a fixed ontology of elements and a fixed set of properties of interest. We consider two classes of properties: symbolic labels for diagram elements and analogical relations among diagram elements. Formally, we can represent the information about labels and relations for diagram elements that is stored in an analogic representation S as a set of first-order models M_S . While a diagram records only those relationships and elements that are known to exist, each of these *diagram models* constitutes a possible completion of the partial information provided by a diagram. For example, the type of U_2 and the owners of U_2 and U_3 are unspecified in the above diagram; a diagram model would fully specify those relations.

Diagram models consist of a set of analogical relations A and a set of label relations L over a universe U . Each member of A is a binary relation $E_s \times E_s$, with $E_s \subset U$ the set of diagram elements; each member of L is a relation $E_s \times E_l$, with $E_l \subset U$ the set of labels. Using the "displayed" format of [2, Section 1.3], we write these models as $\langle U, A, L, E_s, E_l \rangle$.

For the scene described by (1), the diagram elements are $\{V, U_1, U_2, U_3\}$. We choose the label relations $TYPE(u, l)$ and $OWNS(u, l)$, and the analogical relations $BES(u, v)$ (the opening u is next to the opening v) and $INHALL(u, v)$ (opening u is in hall v). The label set contains $\{Closet, Office, Ralph, Paul, Cyril\}$ and possibly other values. The choice of relations and elements is important in determining what information in the analogic structure is abstracted in the hybrid system; here, for example, whether an opening is to the right or left of another opening is apparent from the structure, but not in the models.

A key feature of analogical representations is their capacity to implicitly embody constraints that other representations must make explicit. For example, the map structures embed the following constraints:

- Each opening has at most 2 adjacent openings.
- Objects can have exactly 1 type.
- Individuals can own offices but not closets.
- At most one person can own a given office.

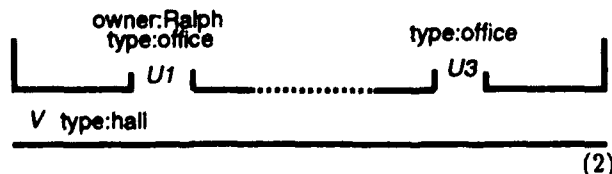
These *diagram constraints* can be built into the representation structures directly or into the operations that manipulate the structures, depending on the given implementation. For example, a bit-map representation of (1) would embed the first constraint directly through its spatial composition; the third constraint would most likely be enforced by operations that manipulate the structure. Either way, diagram constraints are necessarily reflected in diagram models. For instance, all diagram models for (1) can have only one type relation for a given diagram element, due to the second constraint above.

In diagram (1), all objects of relevance (the openings and the hall itself) have been noted and the analogical relations *BES* and *INHALL* are fully determined. Although there is type and ownership information missing, the *structure* of the diagram is complete. Not all diagrams share this completeness. When generating maps from perceptual input, noise or faulty sensors may both cause objects of interest to go undetected and leave analogical relations only partially determined. In such circumstances, we say that the diagram contains *structural uncertainty*. We formalize this notion as follows.

Definition 2.1 (Determined Relation) A set of models M defined over a class of relations $R = \{r_1, \dots, r_m\}$ determines a relation $r_i \in R$ iff every model in M agrees on the extension of r_i .

Definition 2.2 (Structural Uncertainty) A diagram S with models M_S is structurally uncertain iff some analogical relation of the models is undetermined.

The following diagram constitutes a variation on the scene described by (1) in which there is structural uncertainty between U_1 and U_3 . Here, both the *BES* and *INHALL* relations are undetermined. Dashed lines indicate regions of structural uncertainty:



As will be seen, our framework provides the means to apply sentential information about a diagram in order

to both ascertain the composition of areas of structural uncertainty and flesh out the partial characterizations given by the diagram models for the relations in $L \cup A$.

2.2 Sentential Subsystem

The sentential subsystem employs a first-order language

$$\mathcal{L} = (\mathcal{P}_A, \mathcal{P}_L, E_s, E_l, \dots)$$

and a corresponding proof theory. For simplicity, we use the diagram elements E_s and labels E_l as standard names for themselves in \mathcal{L} . The predicates \mathcal{P}_A are interpreted by the analogical relations of the diagram models, and \mathcal{P}_L by the label relations. In addition, there may be other predicates and constants that have an indirect relation to the diagram – for example, the predicate *NBR*(x, y) representing the office-neighbour relationship between two people. This predicate would be related to the diagram predicates $\mathcal{P}_A \cup \mathcal{P}_L$ by an axiom such as

$$\begin{aligned} \forall x, y. \text{NBR}(x, y) \equiv \\ \exists u, v. \text{TYPE}(u, \text{Office}) \wedge \text{TYPE}(v, \text{Office}) \\ \wedge \text{OWNS}(u, x) \wedge \text{OWNS}(v, y) \wedge \text{BES}(u, v). \end{aligned} \quad (3)$$

Similarly, the predicate *RESIDES*(x, h) representing the relationship of an individual x having an office in hallway h would be defined as

$$\begin{aligned} \forall x, h. \text{RESIDES}(x, h) \equiv \\ \exists u. \text{INHALL}(u, h) \wedge \text{TYPE}(u, \text{Office}) \\ \wedge \text{OWNS}(u, x). \end{aligned} \quad (4)$$

We refer to axioms of this sort as *grounding axioms*.

As an example of the expression and use of sentential information relative to diagrams, consider the following statements:

*Paul and Cyril have offices in hall V.
Ralph and Paul are not neighbours.*

Given the grounding axioms (3,4), these statements can be translated into the following formulas of \mathcal{L} :

$$\begin{aligned} \text{RESIDES}(\text{Cyril}, V) \wedge \text{RESIDES}(\text{Paul}, V) \\ \neg \text{NBR}(\text{Ralph}, \text{Paul}). \end{aligned} \quad (5)$$

With respect to diagram (1), the first statement implies that U_2 and U_3 are offices, one each owned by Cyril and Paul. This conclusion follows since $\{U_1, U_2, U_3\}$ constitutes the set of all offices in V and *Ralph* is known to own U_1 . Deduction of this result requires information that is implicit in the diagram's structure, namely that each office can be owned by only one individual. With the second statement, the

only possible configuration of the scene is:



(6)

Sentential information can also be used to reduce structural uncertainty in diagrams: given the sentences *Ralph and Cyril are neighbours* and *Cyril is Paul's only neighbour*, the diagram (6) follows from (2).

2.3 Integration Criteria

In order to determine whether a given integration method behaves in an appropriate fashion, it is necessary to provide a semantic account of the over-all hybrid system.

From a model-theoretic perspective, the merging of a diagram S with a set of sentences T that describe the diagram amounts to restricting the models of S to those that are compatible with T . Compatibility here means that the diagram model can be expanded to a model for T by providing interpretations for the predicate, function, and constant symbols of \mathcal{L} that do not overlap the diagram model.

Definition 2.3 (Restricted Models) Let T be a collection of sentences in \mathcal{L} describing properties of a diagram S and let M_S be the models of S defined for the sublanguage $\langle \mathcal{P}_A, \mathcal{P}_L, E_s, E_l \rangle$ of \mathcal{L} . The restriction of M_S relative to T , written as $M_S(T)$, is the set of models $\langle U, A, L, E_s, E_l \rangle \in M_S$ for which some expansion $\langle U, A, L, E_s, E_l, \dots \rangle$ is a model of T .

The models in $M_S(T)$ characterize the total information content in the hybrid system for the domain modeled by the analogical structures. The challenge is to provide both *derivation rules* for determining formulas of \mathcal{L} that are logically entailed by $M_S(T)$ and *update rules* for modifying S to eliminate diagram models not contained in $M_S(T)$.

In general, the analogical structures may have weaker representational capabilities than is required to capture the information content of $M_S(T)$. Consider the diagram (1) and the sentential theory

$$T_0 = \{ \text{RESIDES}(\text{Cyril}, V), \text{RESIDES}(\text{Paul}, V) \}.$$

These two sources of information jointly imply that U_2 and U_3 are offices, one each owned by *Paul* and *Cyril*; however, it is undetermined as to who owns which one. Every model in $M_S(T_0)$ either has both $\langle U_2, \text{Cyril} \rangle$ and $\langle U_3, \text{Paul} \rangle$ or both $\langle U_2, \text{Paul} \rangle$ and $\langle U_3, \text{Cyril} \rangle$ in its OWNS relation. However, this information cannot be fully manifest in the diagram since it is not definite

about who owns which office and the diagram does not admit disjunctive information about ownership.

Rather than seeking an analogical structure with the models $M_S(T)$, the best that can be attained is a structure that adequately represent $M_S(T)$:

Definition 2.4 (Representational Adequacy)

An analogical structure Q adequately represents a set of diagram models M iff $M \subseteq M_Q$ and there is no other diagram R such that $M \subseteq M_R$ and $M_R \subset M_Q$.

For example, when S is the diagram (1) and T_0 is as defined above, the following diagram adequately represents $M_S(T_0)$:



(7)

This diagram extends (1) to include the information that U_2 is an office but does not include any new information about ownership.

Soundness and completeness for inference in our hybrid system can be defined using the concepts of restricted models and representational adequacy.

Definition 2.5 (Soundness) A diagram update rule is sound iff for diagram S and theory T it generates only diagrams whose model set contains $M_S(T)$. A derivation rule is sound iff it generates only sentences whose model set contains $M_S(T)$.

We will say that a collection of both diagram update and derivation rules is sound precisely when each of its members is sound.

Definition 2.6 (Completeness) A set of derivation and update rules is derivationally complete for S and T iff any valid sentence of $M_S(T)$ can be derived by the sentential subsystem. The set is diagrammatically complete iff it can generate a diagram that adequately represent $M_S(T)$.

3 The Inferential Calculus

The inferential calculus underlying our hybrid framework is defined relative to a class of domain-independent diagram operations. In this section, we describe both the inference rules and operations. We focus exclusively on diagrams without structural uncertainty; diagrams with structural uncertainty are considered in Section 4.

3.1 Diagram Operations

Two classes of diagram operations are required for our inference rules: *reflection* and *extraction* procedures.

Reflection procedures provide a means of inserting information into an analogical structure. For each label predicate $P(u, v)$ we require a reflection procedure $\text{INSERT}.P(u, v)$ such that for $e_1, e_2 \in E_s \cup E_l$, the predicate $P(e_1, e_2)$ holds in all models of the diagram obtained by executing $\text{INSERT}.P(e_1, e_2)$. That is, when applied to a diagram S with model set M_S , $\text{INSERT}.P(e_1, e_2)$ yields a diagram S' with model set

$$M_{S'} = \{m \in M_S \mid m \models P(e_1, e_2)\}.$$

For diagrams without structural uncertainty (the focus of this section), insertion procedures for \mathcal{P}_A are unnecessary.

Extraction procedures provide access to the contents of the analogical structure for use by the sentential subsystem. As noted above, whole-scale translation of the analogical structures into first-order sentences is infeasible. Instead, we wish to provide access to the information in the analogical structures on an *as needed* basis, whereby information is accessed as required for individual deduction steps rather than all at once. The two key types of information stored within diagrams are (1) analogical and label relationships for diagram elements, and (2) closure information about those relationships.

For each diagram predicate $P(u, v)$, we require an extraction procedure $\text{EVAL}.P(u, v)$ for evaluating ground instances relative to the diagram S . These *evaluation* procedures provide the sentential reasoner with information about primitive relationships in the analogical structures. The procedure behaves as follows for $e_1, e_2 \in E_s \cup E_l$:

$$\text{EVAL}.P(e_1, e_2) = \begin{cases} \text{true} & \text{if } M_S \models P(e_1, e_2) \\ \text{false} & \text{if } M_S \models \neg P(e_1, e_2) \\ \text{unknown} & \text{otherwise} \end{cases}$$

Closure information for a diagram S is generated by two classes of procedures. Let $P[x]$ represent an instance of a predicate in $\mathcal{P}_L \cup \mathcal{P}_A$ that contains the single variable x , such as $\text{BES}(x, U_1)$. (For simplicity, we restrict attention here to predicates containing only one variable.) With respect to the diagram S , the procedure $\text{CLOSURE}^+.P[x]$ generates the set of diagram elements that possibly satisfy $P[x]$ (called the *minimal superclosure*) while the procedure $\text{CLOSURE}^-.P[x]$ generates the set of elements that definitely satisfy $P[x]$ (the *maximal subclosure*).

Definition 3.1 (Closures) Let $P[x]$ be a nonground instance of a predicate in $\mathcal{P}_L \cup \mathcal{P}_A$. The minimal superclosure of $P[x]$, denoted by $\text{CLOSURE}^+.P[x]$, and the maximal subclosure of $P[x]$, denoted by

$\text{CLOSURE}^-.P[x]$, are defined with respect to a diagram S as follows:

$$\text{CLOSURE}^+.P[x] = \{e \in E_l \cup E_s \mid m \models P[e] \text{ for some } m \in M_S\}$$

$$\text{CLOSURE}^-.P[x] = \{e \in E_l \cup E_s \mid M_S \models P[e]\}$$

The procedures $\text{CLOSURE}^+.P[x]$ and $\text{CLOSURE}^-.P[x]$ give minimal upper- and maximal lower-bounds, respectively, for the precise set of values that satisfy $P[x]$. This set is fixed for a given diagram only when the relation $P[x]$ is *determined* by the models of S (in the sense of Definition 2.1). In terms of the sentential language \mathcal{L} , determination of $P[x]$ is equivalent to the condition that for $m_1, m_2 \in M_S$:

$$\forall e \in E_s \cup E_l. m_1 \models P[e] \equiv m_2 \models P[e]. \quad (8)$$

When a predicate $P[x]$ is determined by a diagram, the maximal sub- and minimal superclosures are both equal to the exact closure. However, sub- and superclosures are useful sources of diagram information when the predicate is not determined, as will be made apparent in Section 3.2.3.

Note that since we are considering only diagrams without structural uncertainty in this section, all analogical predicates are necessarily determined.

3.2 Inference Rules

The inferential component of the integration framework consists of rules of *evaluation*, *domain enumeration* and *reflection*. Evaluation and domain enumeration utilize information from the diagram as provided by the extraction procedures to derive new sentences describing properties of the diagram. The reflection rule permits the insertion of sentential consequences derived from T into the diagrams using the reflection procedures.

In the definition of the inference rules, we use the notation α_b^c to represent the expression α with all occurrences of the expression b replaced by c .

3.2.1 Reflection

The reflection rule sanctions the transfer of information from the sentential to the analogical subsystem. Let $T \vdash \phi$ represent the deducibility of a sentence ϕ from a set of sentences T using the proof theory of the sentential subsystem.

Definition 3.2 (Reflection Rule) Let T be a sentential theory and S an analogical structure. If $T \vdash R(t_1, \dots, t_k)$ for $t_1, \dots, t_k \in E_s \cup E_l$ and $R \in \mathcal{P}_A \cup \mathcal{P}_L$ then $R(t_1, \dots, t_k)$ can be reflected into S by executing $\text{INSERT}.R(t_1, \dots, t_k)$.

3.2.2 Evaluation

The evaluation rule sanctions replacement of ground instances of a predicate $R \in \mathcal{P}_A \cup \mathcal{P}_L$ by either *true* or *false*, in accordance with the contents of the analogical structure. In the case where the relationship denoted by R is undetermined, the evaluation process has no effect.

Definition 3.3 (Evaluation Rule) Let ϕ be a formula that contains an instance $R(t_1, \dots, t_k)$ of a predicate $R \in \mathcal{P}_A \cup \mathcal{P}_L$. If $\text{EVAL}.R(t_1, \dots, t_k) = \theta$ where $\theta \in \{\text{true}, \text{false}\}$ then evaluation of $R(t_1, \dots, t_k)$ in ϕ yields $\phi_{R(t_1, \dots, t_k)}^\theta$.

3.2.3 Domain Enumeration

The domain enumeration rules allow the elimination of quantifiers in certain cases through the introduction of an appropriate domain of values that covers the relevant instantiations of the quantified variable.

Consider the assertion

$$\exists u. \text{BES}(u, U_2) \wedge \text{OWNS}(u, \text{Paul}) \quad (9)$$

relative to diagram (1). The interpretation of this formula is that the diagram element owned by Paul is located beside U_2 . The conjunct $\text{BES}(u, U_2)$ limits the possibilities for this diagram element: according to (1), the element must be either U_1 or U_3 . As such, the formula $\text{OWNS}(U_1, \text{Paul}) \vee \text{OWNS}(U_3, \text{Paul})$ follows from (9). Similarly, the universally quantified formula

$$\forall u. \text{INHALL}(u, V) \supset \text{TYPE}(u, \text{Office}) \quad (10)$$

can be viewed as a statement about the predicate $\text{TYPE}(u, \text{Office})$, with $\text{INHALL}(u, V)$ serving as a filter on the set of relevant instantiations of the quantified variable. According to the diagram (1), the only values that satisfy $\text{INHALL}(u, V)$ are $\{U_1, U_2, U_3\}$ (i.e., the exact closure of $\text{INHALL}(u, V)$ is $\{U_1, U_2, U_3\}$). Thus, the conjunction

$$\bigwedge_{d \in \{U_1, U_2, U_3\}} \text{TYPE}(d, \text{Office})$$

is equivalent to (10) with respect to models for diagram (1).

We refer to the technique used above for applying closure information to eliminate quantifiers as *domain enumeration*. Domain enumeration does not apply to all predicate instances containing a quantified variable. The formula $\exists u. \neg \text{BES}(u, U_1) \wedge \text{TYPE}(u, \text{Closet})$ illustrates this point. In this case, the exact closure for $\text{BES}(u, U_1)$ is not an appropriate restriction of the terms of \mathcal{L} ; elimination of the existential quantifier using the exact closure would lead to unsound conclusions.

For existential quantifiers, the domain used in domain enumeration must include all bindings for which the

embedded formula (e.g., $\text{BES}(x, U_2) \wedge \text{OWNS}(x, \text{Paul})$ in (9)) may have truth value *true*; this guarantees that all relevant instantiations of the quantified variable are covered. For universal quantifiers, the domain should exclude values for which the embedded formula is already determined to have truth value *true*. We call a predicate instance whose exact closure satisfies these conditions *focus expressions* for the given quantified formula. In essence, a focus expression prunes from consideration those bindings of a given quantified variable that do not provide useful information.

To formalize the concept of focus expressions, we introduce definitions for the *polarity* and *definiteness* of predicate instances in a formula.

Definition 3.4 (Polarity) An instance of a predicate in a formula ϕ is called *positive* if the instance maps to an unnegated literal in the conjunctive normal form of ϕ and is called *negative* otherwise.

Definition 3.5 (Definiteness) An instance of a predicate in a formula ϕ is called *definite* if the instance maps to a literal in a clause of length one in the conjunctive normal form of ϕ and is called *indefinite* otherwise.

We will combine the notions of polarity and definiteness, referring to individual instances as *negative indefinite* or *positive definite* as appropriate. The expression $\text{INHALL}(u, V)$ is a negative indefinite instance in (10) and a positive definite instance in

$$\exists u. \text{INHALL}(u, V) \wedge \text{TYPE}(u, \text{Closet})$$

Definition 3.6 (Focus Expression) If ψ is a quantified formula (either $\forall z. \alpha$ or $\exists z. \alpha$) containing a predicate instance $P[z]$ then $P[z]$ is a focus expression for ψ iff either

- ψ has the form $\forall z. \alpha$ and the occurrence of $P[z]$ is negative indefinite, OR
- ψ has the form $\exists z. \alpha$ and the occurrence of $P[z]$ is positive definite.

The domain enumeration rule is formally defined as follows.

Definition 3.7 (Domain Enumeration Rule) If ψ is a quantified expression, either $\exists z. \alpha$ or $\forall z. \alpha$, containing a focus expression $\Phi[z]$ with maximal subclosure D^- and minimal superclosure D^+ then domain enumeration for ψ and $\Phi[z]$ yields:

$$\bigvee_{d \in D^+} (\alpha_z^d) \quad \text{if } \psi \text{ is } \exists z. \alpha$$

$$\bigwedge_{d \in D^-} (\alpha_z^d)^{\text{true}}_{\Phi[d]} \quad \text{if } \psi \text{ is } \forall z. \alpha$$

Note that when applying domain enumeration to a universally quantified formula $\forall z. \alpha[z]$, the embedded formula $\alpha[z]$ need not be fully retained. Instead, the simplification of $\alpha[z]$ in which the focus expression is replaced by *true* suffices, since the focus expression has truth value *true* for all terms in its maximal subclosure. For example, $INHALL(u, V) \supset TYPE(u, Office)$ can be reduced to the expression $TYPE(u, Office)$. Focus expressions must be retained for existentially quantified formulas: by definition the maximal superclosure may contain terms that are not in the exact closure (and hence do not satisfy the focus expression).

Domain enumeration could be extended to make use of nonatomic focus expressions. For example, the conjunction $BES(x, U_5) \wedge TYPE(x, Office)$ could serve as a focus expression in the formula

$$\forall x. BES(x, U_5) \wedge TYPE(x, Office) \supset OWNS(x, Eva).$$

The intersection of the maximal subclosures for the individual conjuncts would serve as a more restricted (and hence more useful) domain for the universally quantified variable x . Similarly, the disjunction $BES(x, U_4) \vee BES(x, U_6)$ could be employed as a focus expression in

$$\exists x. (BES(x, U_4) \vee BES(x, U_6)) \wedge OWNS(x, Ann).$$

The appropriate domain in this case would be the union of the minimal superclosures for each disjunct. Straightforward extensions of Definitions 3.6–3.7 support this generalization; we forego their technical statement in this paper.

3.3 Example

We illustrate the workings of our integration rules by applying them to the scenario presented in Section 2.2, namely diagram (1) with sentential theory

$$T_0 = \{\neg NBR(Ralph, Paul), RESIDES(Cyril, V), RESIDES(Paul, V)\}.$$

A derivation schematic, including both diagrams and sentences, is provided in Figure 1.

Consider first the given formula $RESIDES(Paul, V)$. Rewriting using definition (4) yields formula S2 in the figure. The predicate $INHALL(u, V)$ is a focus expression in S2 and its exact closure in diagram (a) is $\{U_1, U_2, U_3\}$. Domain enumeration using this focus expression yields formula S3. Diagram (a) contains the information that U_1 and U_3 are offices, thus $TYPE(U_1, Office)$ and $TYPE(U_3, Office)$ in S3 can be replaced by *true* using the evaluation rule. In addition, since the diagram indicates that *Ralph* owns U_1 , evaluation can be used to rewrite $OWNS(Paul, U_1)$ to *false*. This evaluation step exploits the diagram constraint that ownership is unique. The evaluations combined with tautological simplification produce formula S4.

Expansion of the given fact $\neg NBR(Ralph, Paul)$ using definition (3) gives formula S6 in the figure. This formula contains the focus expression $OWNS(Ralph, x)$

whose exact closure in diagram (a) is $\{U_1\}$; domain enumeration is applied to produce formula S7. Evaluation of the expression $TYPE(U_1, Office)$ with respect to diagram (a) leads to formula S8, which contains the focus expression $BES(U_1, y)$ whose exact closure is $\{U_2\}$. Domain enumeration for $BES(U_1, y)$ yields formula S9, which along with S4 entails $OWNS(Paul, U_3)$. Application of the reflection rule to this atom yields the second diagram (b).

The formula S12 is obtained from the given formula $RESIDES(Cyril, V)$ by applying the same steps used from S1 to S4 above. The ownership of U_3 was undetermined in the original diagram; however, $OWNS(Cyril, U_3)$ is necessarily *false* since the new diagram (b) indicates that the owner of U_3 is *Paul*. Evaluation can be applied to the formula S12 using diagram (b) to derive S13. Note that this last deduction could not be made from S12 and the sentence $OWNS(Paul, U_3)$ alone; again we need the diagram constraint that ownership is unique. Finally, the contents of this last formula can be reflected to produce the diagram (c), which adequately represents $M_S(T_0)$.

4 Structural Uncertainty

Consider the diagram



containing a region of structural uncertainty between elements U_1 and U_3 . This diagram has models in which there are zero, one, two, etc, diagram elements in the uncertain area. Without further information, there is no way to determine which of these models corresponds to the actual situation that the diagram is intended to represent.

Accounting for structural uncertainty requires a slight generalization of the inferential calculus presented in Section 3. First of all, reflection operators must be provided for the analogical predicates \mathcal{P}_A so that diagrams can be modified to incorporate new analogical relations determined by the sentential subsystem. The definitions of the remaining diagram operations and the various inference rules remain unaltered but the definition of minimal superclosure requires elaboration. Because the minimal superclosure of a predicate must include all possible values for which the predicate holds, it is necessary to consider whether elements inserted in structurally indeterminate regions could satisfy the given predicate. In contrast, the definition of maximal subclosure is not affected by structural uncertainty since diagram elements that do not appear in all models are not included in the maximal

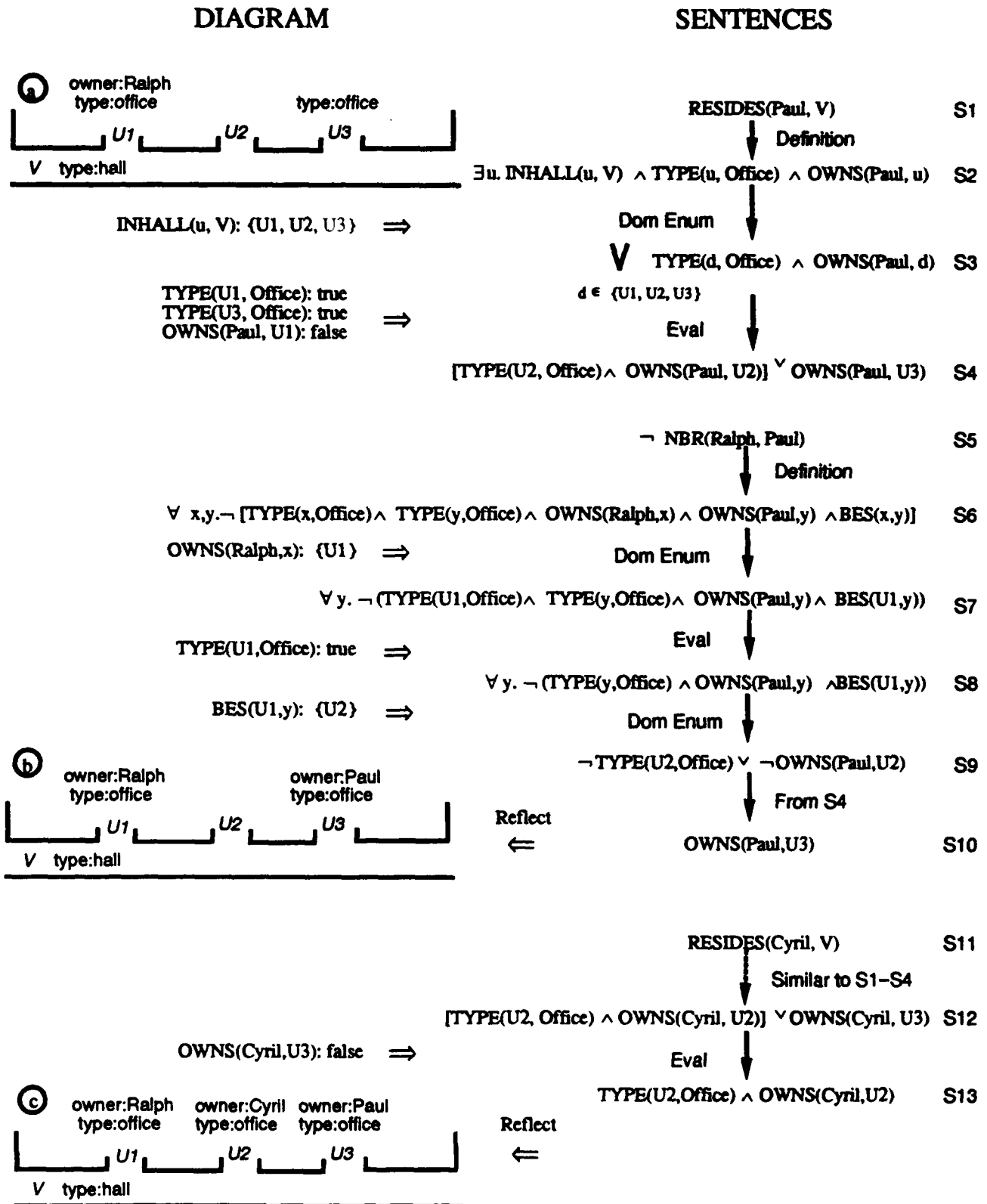


Figure 1: An Example Derivation

subclosure.

4.1 Minimal Superclosures with Introduced Names

To account for structural uncertainty in diagrams, we exploit a technique of the natural deduction calculus for dealing with existential elimination. With this calculus, the existential quantifier of a formula $\exists x.\phi[x]$ can be eliminated by introducing a *new* individual constant c for x , yielding $\phi[c]$. The justification for the introduction is that, since c does not appear elsewhere in the proof, it can refer to an arbitrary individual.

We employ the same principle in formulating minimal superclosures for diagrams containing structural uncertainty. Consider diagram (11) relative to the sentence $\exists x. BES(U_1, x)$. It could be that U_3 is next to U_1 , or that there is an intervening element I_1 situated to the right of U_1 . The minimal superclosure must take both of these cases into account, yielding the set $\{U_3, I_1\}$.

Employing names for hypothesized individuals introduces a complication to the diagrams, since we have heretofore assumed that all elements were "standardized apart," receiving different names if and only if they were distinct. This is not the case with hypothesized individuals; for example, if we were later to perform domain enumeration on the sentence $\exists y. BES(U_3, y)$, introducing the name I_2 , it could be the case that both I_1 and I_2 refer to the same individual. To account for naming and identity, we assume that the diagram keeps track of introduced names and their possible referents.¹

When the exact closure of an analogical predicate $P[x]$ is *determined* by the diagram models (i.e., condition (8) is satisfied), the minimal superclosure reduces to the exact closure. Otherwise, the minimal superclosure consists of those diagram elements that satisfy the predicate in any diagram model, along with an introduced name for a hypothesized element. Even though multiple elements can appear in regions of uncertainty and there may be more than one such region in a diagram, only one introduced element is required for the minimal superclosure. Restriction to one such element is possible because the purpose of domain enumeration is to identify a solitary element satisfying the matrix of the existentially quantified formula.

Definition 4.1 (Minimal Superclosure for \mathcal{P}_A)
Let $P[x]$ represent an instance of a predicate in \mathcal{P}_A that contains the single variable x . The minimal superclosure of $P[x]$, denoted by $CLOSURE^+.P[x]$, is defined for a diagram S as

$$\{e \in E_i \cup E_s \mid M_S \models P[e]\}$$

¹In other words, diagram operations must track equalities and inequalities for introduced names.

when $P[x]$ is determined by M_S , otherwise

$$\{e \in E_i \cup E_s \mid m \models P[e] \text{ for some } m \in M_S\} \cup \{I_k\}$$

where I_k is an introduced name.

The minimal superclosure for $BES(U_1, x)$ relative to diagram (11) is $\{U_3, I_1\}$, where I_1 is an introduced name.

The new definition of minimal superclosure is used as before in domain enumeration except that when an element name I_k is introduced for a minimal superclosure, the diagram is modified to include inequalities between I_k and all current diagram elements. The referent of an introduced I_k may be determined by future sentential reasoning steps, possibly leading to the insertion of a new diagram element via the reflection rule. Such a situation arises in the example presented below.

4.2 Example: Structural Uncertainty

Consider diagram (11) and the theory

$$T_1 = \{NBR(Ralph, Cyril), NBR(Paul, Cyril)\}.$$

Every model in $M_S(T_1)$ has exactly one element between U_1 and U_3 , with this element being labeled as the office of Cyril. We show how a new diagram that reflects this information can be generated using the integration calculus.

Applying the same steps used to derive S8 from $\neg NBR(Ralph, Paul)$ in Figure 1, we generate the following pair of formulas from T_1 :

$$\exists x. BES(U_1, x) \wedge TYPE(x, Office) \wedge OWNS(Cyril, x) \quad (12)$$

$$\exists x. BES(U_3, x) \wedge TYPE(x, Office) \wedge OWNS(Cyril, x). \quad (13)$$

As noted above, the minimal superclosure for $BES(U_1, x)$ relative to diagram (11) is $\{U_3, I_1\}$. Domain enumeration for the focus expression $BES(U_1, x)$ in (12) yields

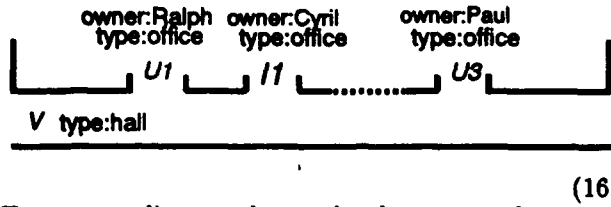
$$\begin{aligned} & BES(U_1, U_3) \wedge TYPE(U_3, Office) \wedge OWNS(Cyril, U_3) \\ & \vee \\ & BES(U_1, I_1) \wedge TYPE(I_1, Office) \wedge OWNS(Cyril, I_1) \end{aligned} \quad (14)$$

along with the naming constraints $U_3 \neq I_1$ and $U_1 \neq I_1$. Evaluation of $OWNS(Cyril, U_3)$ with respect to the diagram returns *false* (since ownership is unique), thus (14) simplifies to

$$BES(U_1, I_1) \wedge TYPE(I_1, Office) \wedge OWNS(Cyril, I_1). \quad (15)$$

The conjuncts in this formula can be reflected to pro-

duce the new diagram



Here, a new diagram element has been created to serve as the referent of the introduced name I_1 .

In diagram (16), the expression $BES(x, U_3)$ has minimal superclosure $\{I_1, I_2\}$, for some introduced name I_2 . Domain enumeration for $BES(x, U_3)$ in (13) gives

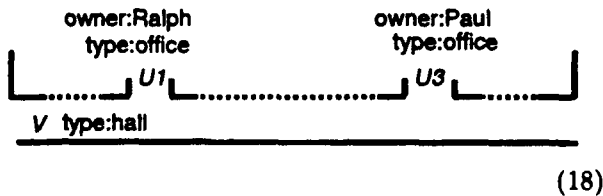
$$\begin{aligned} & BES(U_3, I_1) \wedge TYPE(I_1, Office) \wedge OWNS(Cyril, I_1) \\ & \vee \\ & BES(U_3, I_2) \wedge TYPE(I_2, Office) \wedge OWNS(Cyril, I_2) \end{aligned} \quad (17)$$

along with the inequalities $I_2 \neq U_1$, $I_2 \neq U_3$, and $I_2 \neq I_1$.

Since the current diagram indicates that Cyril owns I_1 (and cannot own I_2 since $I_2 \neq I_1$), the evaluation rule can be applied to eliminate the second disjunct of (17). Further evaluations lead to the formula $BES(U_3, I_1)$, thus establishing that the introduced name I_2 does not refer to a realizable diagram element. Reflection of this last atom yields the diagram (6), which adequately represents $M_S(T_1)$.

4.3 A Troublesome Example

The new definition of minimal superclosure does not always lead to an adequate representation of the restricted class of diagram models. Suppose that we relate sentences (12,13) to the following diagram D:



The models $M_D(T_1)$ are adequately represented by a diagram similar to (6) but with regions of uncertainty to the left of U_1 and to the right of U_3 . As we will show, our integration calculus cannot produce this diagram.

The minimal superclosure for $BES(U_1, x)$ is again $\{U_3, I_1\}$ as was the case in the previous example (provided we reuse I_1 as the introduced name) and we can similarly derive the formula (15) from (12). However, since I_1 could be located on either side of U_1 we cannot insert a new element into the diagram as the referent of I_1 . This inability to situate I_1 leads to a different minimal superclosure for $BES(U_3, x)$, namely $\{U_1, I_2\}$ (again we reuse the same introduced name

from the previous example). Domain enumeration for $BES(x, U_3)$ in (13) now gives

$$\begin{aligned} & BES(U_3, U_1) \wedge TYPE(U_1, Office) \wedge OWNS(Cyril, U_1) \\ & \vee \\ & BES(U_3, I_2) \wedge TYPE(I_2, Office) \wedge OWNS(Cyril, I_2) \end{aligned}$$

with the inequalities $I_2 \neq U_1$ and $I_2 \neq U_3$. Note that in this case, the inequality $I_2 \neq I_1$ is not added since I_1 does not refer to a current diagram element.

At this point, no modifications to the diagram are possible, and no further derivations by the sentential subsystem lead to any reflections back to the diagram. While the atoms $BES(U_1, I_1)$, $BES(U_3, I_2)$, $OWNS(Cyril, I_1)$ and $OWNS(Cyril, I_2)$ are all derivable, they have no effect on the diagram individually. In combination though, they constrain $I_1 = I_2$ to be the unique office situated between U_1 and U_2 . Generating a diagram that adequately represents $M_D(T_1)$ requires reasoning by cases about the possible locations of introduced elements and is beyond the scope of the integration calculus presented in this paper.

5 Properties of the Framework

The integration framework satisfies the following properties.

Proposition 5.1 (Soundness) *Reflection, evaluation and domain enumeration are sound.*

An inference rule that derives a formula ψ from a given formula ϕ is *equivalence-preserving* with respect to a class of models M when $M \models \phi \equiv \psi$.

Proposition 5.2 (Equivalence) *Domain enumeration using exact closures for a diagram S is an equivalence-preserving inference rule with respect to the models of S .*

The proofs of the propositions are quite straightforward; we defer them to a longer version of the paper.

The integration rules are neither derivationally nor diagrammatically complete. The central problem is that the rules focus on properties of individuals and their relationships with other individuals, failing to account for embedded diagram constraints. Diagram constraints are certainly made use of at times. In going from S_3 to S_4 in Figure 1, the unique ownership constraint made it possible to conclude that $OWNS(Paul, U_1)$ had truth-value *false*, given that $OWNS(Ralph, U_1)$ held in diagram (a). However, it is impossible to directly reason with the diagram constraints.

In the remainder of this section we briefly consider the issue of completeness for propositional sentential theories. We note that the discussion is also of relevance to those first-order theories that can be reduced to

propositional theories through appropriate use of the domain enumeration rule.

Derivational completeness demands the derivability in the sentential subsystem of any sentence valid in $M_S(T)$. Suppose we pick a propositionally-complete refutation strategy for the sentential subsystem. Is the resulting system complete? The answer is "no." Consider the following set of statements:

$$\begin{aligned} \text{OWNS}(\text{Paul}, U_1) &\vee \text{OWNS}(\text{Paul}, U_2) \\ \text{OWNS}(\text{Ralph}, U_1) &\vee \text{OWNS}(\text{Ralph}, U_2) \\ \text{OWNS}(\text{Cyril}, U_1) &\vee \text{OWNS}(\text{Cyril}, U_2) \end{aligned}$$

Given the embedded property of unique ownership, these sentences are inconsistent with respect to any diagram in our class of maps. Even so, it is not always possible to derive the empty clause. In particular, no refutation is possible given a variation of diagram (1) in which all ownership information has been removed. Because the uniqueness constraint on ownership is embedded in the representations and operations of the analogical structures, the integration rules provide no means of relating this constraint to the sentences above.

Derivational completeness can be attained by extending the evaluation rule to sets of literals. Define:

$$\text{EVAL}^*(l_1, \dots, l_n) = \begin{cases} \text{inconsis} & \text{if } M_S \models \neg(l_1 \wedge \dots \wedge l_n) \\ \text{unknown} & \text{otherwise} \end{cases}$$

Using this evaluation procedure, we can apply total narrow theory resolution [25] as a refutation-complete derivational procedure. This procedure is a variant of hyperresolution in which a set of literals, one from each clause of the resolution, are tested for consistency against the diagram; if they are inconsistent, the result of the resolution is a clause consisting of a disjunction of the remainders of each resolved clause.²

Diagram completeness is generally harder to achieve than derivational completeness because it requires the sentential subsystem to be complete for atomic consequence-finding. Consider the theory:

$$T = \{\text{OWNS}(\text{Paul}, U_3) \vee \text{OWNS}(\text{Cyril}, U_3)\}$$

relative to diagram (1). Given the embedded diagram constraint that ownership applies only to offices, it is possible to derive $\text{TYPE}(U_3, \text{Office})$ by refutation. But $\text{TYPE}(U_3, \text{Office})$ cannot be derived as a consequence of the diagram and T . Arriving at this conclusion requires a form of reasoning by cases that our framework

²Although we can achieve derivational completeness using theory resolution in a refutation system, in practice we might not want to use theory resolution directly, since it is a multiple-clause inference rule. It would be more efficient to consider a variation on the Davis-Putnam method in which branches are closed when they contain a set of atoms that are inconsistent according to the diagram.

does not currently support. As noted above, reasoning by cases is also required to overcome the form of diagram incompleteness that arose in Section 4.3. This capability presents an interesting avenue for further research.

6 Summary

We have described a domain-independent formal framework for integrating sentential and analogical representations. We illustrated the workings of the framework for the application of reasoning sententially to extend the information content of maps, both with and without structural uncertainty. The integration rules of the framework are sound as well as derivationally complete in the propositional case.

The integration framework has been implemented on top of the KLAUS automated deduction system [26] using the method of *universal attachment* [14, 15] to formulate the integration rules. The system has been successfully applied to problems involving reasoning with maps, including the examples presented in this paper. Much work remains to be done on control issues for the implementation. In particular, the ordering of domain enumerations can greatly impact the efficiency of reasoning.

There has been a resurgence of interest in computational models for diagrammatic/visual reasoning during the past few years. Most similar in nature to our research is the work on Hyperproof [1]. The Hyperproof system combines sentential reasoning with diagrammatic representations of a chessboard containing blocks. In Hyperproof, much of the complexity underlying the integration of diagrammatic and sentential information is implicit in the system; in contrast, our research has sought to provide a domain-independent inferential framework in which all aspects of integration are made explicit. A second difference relates to control: the inference rules presented here automatically combine sentential and diagrammatic reasoning, while Hyperproof requires user interaction to guide the reasoning process.

The work of [17] presents a computational model for reasoning with diagrammatic representations but emphasizes the emulation of human reasoning about visualization. *Computational imagery* [19] defines a representational framework for reasoning with both visual and spatial information but does not address the connection to deductive inference.

Our work also overlaps to a certain extent with research in the hybrid reasoning community [4, 3]. Other than our specialization of hybrid reasoning to analogical representations, the main difference between the material presented there and the hybrid framework presented here is the latter's emphasis on reflecting derived information back into analogical structures.

Acknowledgements

This research was supported by the Office of Naval Research under Contract N00014-89-C-0095.

References

- [1] J. Barwise and J. Etchemendy. Visual information and valid reasoning. In W. Zimmerman, editor, *Visualization in Mathematics*. Mathematical Association of America, Washington, DC, 1990.
- [2] C. Chang and H. J. Keisler. *Model Theory*. Elsevier Press, New York, 1977.
- [3] A. M. Frisch, editor. *Proceedings of the AAAI 1991 Fall Symposium on Principles of Hybrid Reasoning*, 1991.
- [4] A. M. Frisch and R. B. Scherl. A bibliography on hybrid reasoning. *AI Magazine*, 11(5), 1991.
- [5] B. V. Funt. Problem-solving with diagrammatic representations. *Artificial Intelligence*, 13, 1980.
- [6] G. W. Furnas. Formal models for imaginal deduction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 662-669. Lawrence Erlbaum, 1990.
- [7] F. Gardin and B. Meltzer. Analogical representations of naive physics. *Artificial Intelligence*, 38:139-159, 1989.
- [8] H. Gelernter. Realization of a geometry-theorem proving machine. In E. A. Feigenbaum and J. Feldman, editors, *Computers and Thought*. McGraw-Hill, New York, 1963.
- [9] P. J. Hayes. Some problems and non-problems in representation theory. In *Proceedings of the AISB Summer Conference*, pages 63-79, University of Sussex, 1974.
- [10] P. N. Johnson-Laird. Mental models in cognitive science. *Cognitive Science*, 4:71-115, 1980.
- [11] S. M. Kosslyn. *Image and Mind*. Harvard University Press, Cambridge, MA, 1980.
- [12] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11:65-99, 1987.
- [13] K. L. Myers. *Universal Attachment: An Integration Method for Logic Hybrids*. PhD thesis, Stanford University, 1991.
- [14] K. L. Myers. Universal attachment: An integration method for logic hybrids. In J. A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*. Morgan Kaufmann, 1991.
- [15] K. L. Myers. Hybrid reasoning using universal attachment. *Artificial Intelligence*, 1992. To appear.
- [16] K. L. Myers and K. Konolige. Integrating analogical and sentential reasoning for perception. In *Proceedings of the AAAI Spring Symposium on Reasoning with Diagrammatic Representations*, 1992.
- [17] N. Narayanan and B. Chandrasekaran. Reasoning visually about spatial interactions. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 1991.
- [18] E. P. Novak, Jr. and W. C. Bulko. Understanding natural language with diagrams. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 465-470, 1990.
- [19] D. Papadias and J. I. Glasgow. A knowledge representation scheme for computational imagery. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, 1991.
- [20] S.-J. Shin. An information-theoretic analysis of valid reasoning with Venn diagrams. In J. Barwise, J. M. Gawron, G. Plotkin, and S. Tutiya, editors, *Situation Theory and its Applications*, volume 2. 1991.
- [21] S.-J. Shin. *Valid Reasoning and Visual Representation*. PhD thesis, Dept. of Philosophy, Stanford University, 1991.
- [22] A. Sloman. Interactions between philosophy and AI. *Artificial Intelligence*, 2, 1971.
- [23] A. Sloman. Afterthoughts on analogical representation. In *Proceedings of Theoretical Issues in Natural Language Processing*, 1975.
- [24] K. Stenning and J. Oberlander. Spatial containment and set membership. In J. Barnden and K. Holyoak, editors, *Analogical Connections*. 1992.
- [25] M. E. Stickel. Automated deduction by theory resolution. *Journal of Automated Reasoning*, 1(4), 1985.
- [26] M. E. Stickel. The KLAUS automated deduction system. In *Proceedings of the Ninth International Conference on Automated Deduction*, 1988.

Integrating Analogical and Sentential Reasoning for Perception

Karen L. Myers Kurt Konolige

Artificial Intelligence Center

SRI International

333 Ravenswood Ave.

Menlo Park, CA 94025

Abstract

Many diverse sources can contribute to the successful interpretation of sensory input. One fundamental problem for perception is integrating these sources into the interpretation process. We present a hybrid methodology for perception that addresses this integration problem. The approach integrates special-purpose analogical representations used to store partially interpreted sensor data with a general-purpose sentential language employed to represent more cognitively based information about a domain. The paper describes a formal apparatus for unifying the sentential and analogical representations as well as inference mechanisms for translating between the two subsystems. We exhibit the utility of the framework for the task of integrating contingent information into maps.

1 Introduction

High-level symbolic representations of information play an important part in perceptually grounded intelligent systems. The need to connect sensory input to high-level representation structures is apparent for systems that perform cognitive operations, such as mobile robots that formulate run-time plans. Traditionally, attention has focused on the flow of information from perceptual input to a high-level symbolic language; however, the transfer of information in the opposite direction can also be valuable. In particular, many important perceptual tasks can be executed more robustly when symbolic reasoning is incorporated into the process of perception itself.

To date, most perceptual interpretation systems have relied exclusively on special-purpose representations and algorithms that embed model information for a class of perceptual tasks. Although these representations and algorithms are essential to timely interpretation, we argue that the inclusion of more general-purpose representation and reasoning mechanisms can greatly extend the capabilities of the system. In support of this claim, consider the role of *contingent information* [7] in perception. Contingent information describes properties that hold for a particular situation or context rather than in the general case, such as the fact that the stairwell is beside the elevator in a given building or that a given office belongs to a certain individual. Contingent information is essential for deter-

mining characteristics of an environment that cannot be directly perceived by the mobile robot. For example, a robot cannot determine the owner of a particular office using range-finding sensors. The ability to incorporate contingent information into the perceptual interpretation process would make it possible to eliminate such gaps in the robot's representation of its environment. The difficulty, however, is that contingent information, and other types of information that can contribute to the interpretation process, may be expressed as graphs, diagrams, logical formulae or in some other format that differs substantially from the geometrical representations prevalent in sensory analysis.

This paper presents a hybrid framework for perception that supports the incorporation of diverse sources of information into the perceptual interpretation process. Special-purpose analogical representations serve as the primary repository for geometrical interpretations of sensor data, while a general-purpose logical language represents more cognitively based information about the domain. We call this latter *sentential information* because it is expressed in the form of sentences in a language. Successful perception involves many layers of interpretation and hence a corresponding hierarchy of representation structures [2]. Generally speaking, we can subdivide these layers into three sections: *sensor data* is refined into an *image-level* representation, which is further interpreted to provide a *scene-level* description [7]. Although sentential reasoning can be put to good use throughout, we are interested primarily in applying sentential reasoning to improve a scene-level interpretation of sensed information.

The fundamental technical challenge in building the hybrid framework is bridging the gap between sentential and scene-level representations. We present a formal apparatus for connecting sentential and analogical representations along with efficient inference mechanisms for translating between the two. In contrast to previous work in the hybrid systems community, our framework supports the reflection of information derived through sentential reasoning back to the analogical structures. For concreteness, we consider the specific problem of employing contingent information in the task of map-learning. Many of the results, however, apply to arbitrary hybrid architectures that link analogical and sentential subsystems.

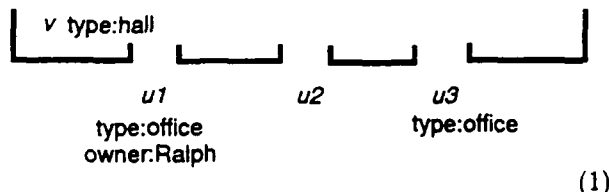
2 The Hybrid Framework

Integration should not be achieved in the hybrid system by simply translating the contents of one representation language into the other. Analogical and sentential structures are effective for representing different types of information. Analogical structures provide a convenient means of representing closure information implicitly while also permitting direct access to analogical properties; however, such structures lack the expressive power of formal logic. Since sentential theories are a more general representational technology, it is tempting to translate analogical structures into first-order sentences *en masse*. But this strategy would compromise the efficiency of the representation system since the specialized inference mechanisms for the analogical structures are replaced by general-purpose deductive methods. Our strategy is to build separate analogical and sentential subsystems along with inference rules for translating information between them.

2.1 Analogical Subsystem

The details of the analogical subcomponent will vary for different applications. Our formal framework isolates the integration methods from the specifics of any particular application through the use of an abstract characterization of the information stored in the analogical system. Since we are interested primarily in analogical structures employed for map-learning in an office environment, we employ examples from that domain throughout this document.

A typical scene-level description of one side of a hallway stored in the representation structures of a robot might contain the following information:¹



The constants V and U_i are symbolic names assigned to the hallway and the three openings on it in the given scene. These objects and the relationships among them are identified by the robot's perceptual interpretation mechanism, which detects relevant geometric properties and segments sensory input into meaningful units (e.g., groups line segments and intersegment spaces into objects such as corridors and significant openings). We use the term *scene element* for such objects. Prior knowledge about the scene was used to determine the remainder of the information in this diagram, namely that certain U_i are offices and that the leftmost office belongs to Ralph.

For any particular class of applications, there will be a fixed ontology of scene elements and a fixed set of

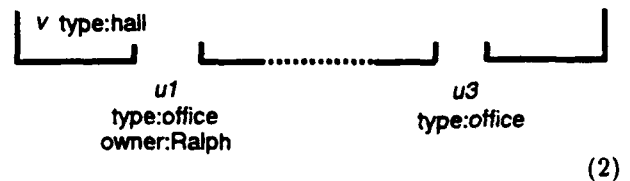
¹The robot would most likely represent this diagram internally using graph-theoretic structures (e.g., nodes representing portals and links between nodes representing the adjacency of portals, along with a higher-level graph for connectivity of hallways).

properties of interest. We consider two classes of properties: assignment of labels to scene elements and analogical relations among scene elements. Formally, we can represent the information about labels and relations for scene elements that is stored in an analogic representation S as a set of first-order models M_S . While a scene structure records only those relationships and elements that are known to exist, each of these *scene models* constitutes a possible completion of the partial information provided by a scene structure. For example, the type of U_2 or the owners of U_2 and U_3 are unspecified in the above diagram; a scene model would fully specify those relations.

Scene models consist of a set of analogical relations A and a set of label relations L over a universe U . Each member of A is a binary relation $E_s \times E_s$, with $E_s \subset U$ a distinguished set of scene elements; and each member of L is a relation $E_s \times E_l$, with $E_l \subset U$ a distinguished set of label elements. Using the "displayed" format of [3, Section 1.3], we write these models as $\langle U, A, L, E_s, E_l \rangle$. Scene models are used as an analytic tool for characterizing the semantic content of an analogic structure; all computations are done on the structure itself.

For the scene described by (1), the scene elements are $\{V, U_1, U_2, U_3\}$. We choose the label relations *TYPE* and *OWNS*, and the scene relations *BES*(u, v) (the opening u is next to the opening v) and *INHALL*(u, v) (opening u is in hall v). The label set consists of $\{\text{Closet, Office, Ralph, Paul, Cyril}\}$. The choice of relations and elements is important in determining what information in the analogic structure is abstracted in the hybrid system; here, for example, whether an opening is to the right or left of another opening is apparent from the structure, but not in the models.

In the example, the entire hallway has been fully perceived. Thus, all objects of relevance (here, the openings and the hall itself) have been detected and the analogical relations *BES* and *INHALL* are fully determined. More generally, noise or faulty sensors may both cause objects of interest to go undetected or uninterpreted and leave analogical relations only partially determined. In such circumstances, we say that the scene structure contains *perceptual uncertainty*. For instance, the following diagram represents a map of the above scene for which the sensor input between U_1 and U_3 could not be interpreted; dashed lines indicate regions of perceptual uncertainty:



The integration problem that we address is to use contingent information about the scene in order to both ascertain the composition of areas of perceptual uncertainty and flesh out the partial characterizations given by the scene models for the relations in L and A .

2.2 Sentential Subsystem

The sentential subsystem employs a first-order language $\mathcal{L} = \langle \mathcal{P}_A, \mathcal{P}_L, E_s, E_l, \dots \rangle$ for expressing contingent information about a scene. For simplicity we have used the scene elements E_s and label elements E_l as names for themselves in \mathcal{L} . The predicates \mathcal{P}_A are interpreted by the analogical relations of the scene models, and \mathcal{P}_L by the label relations. In addition, there may be other predicates and constants that have an indirect relation to the scene – for example, the predicate $NBR(x, y)$ representing the office-neighbour relationship between two people. This predicate would be related to the scene predicates by an axiom such as

$$\begin{aligned} \forall x, y. NBR(x, y) \equiv \\ \exists u, v. TYPE(u, Office) \wedge TYPE(v, Office) \\ \wedge OWNS(u, x) \wedge OWNS(v, y) \wedge BES(u, v) \end{aligned} \quad (3)$$

Similarly, the predicate $RESIDES(x, h)$ representing the relationship of an individual x having an office in hallway h would be defined as

$$\begin{aligned} \forall x, h. RESIDES(x, h) \equiv \\ \exists u. INHALL(u, h) \wedge TYPE(u, Office) \\ \wedge OWNS(x, u) \end{aligned} \quad (4)$$

We refer to axioms of this sort as *perceptual grounding axioms*.

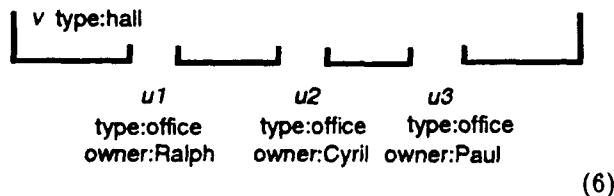
As an example of the expression and use of contingent information relative to diagrams, consider the following statements:

Paul and Cyril have offices in hall V.
Ralph and Paul are not neighbours.

Given the grounding axioms (3,4), these statements can be expressed in \mathcal{L} as

$$\begin{aligned} RESIDES(Cyril, V) \wedge RESIDES(Paul, V) \\ \wedge \neg NBR(Ralph, Paul) \end{aligned} \quad (5)$$

With respect to the scene diagram (1), the first statement implies that U_2 and U_3 are offices, one each owned by Cyril and Paul. With the second statement, the only possible configuration of the scene is the one given below:



When the sentential facts are applied to the scene (2) containing perceptual uncertainty, no updates to the diagram are possible. However, given the sentences *Ralph and Cyril are neighbours* and *Cyril is Paul's only neighbour*, the scene description (6) follows.

2.3 Integration Criteria

We now turn to the problem of characterizing the semantic content of our hybrid system. This characterization will determine what inference mechanisms are

appropriate for combining analogical and sentential information.

Let T be a theory of \mathcal{L} expressing contingent knowledge for a scene and let M_S be the scene models defined for the sublanguage $\langle \mathcal{P}_A, \mathcal{P}_L, E_s, E_l \rangle$ of \mathcal{L} . From a model-theoretic perspective, the incorporation of T into the scene structure S eliminates from M_S those models that are not *compatible* with T . A model $m = \langle U, A, L, E_s, E_l \rangle \in M_S$ is compatible with T if some expansion $m' = \langle U, A, L, E_s, E_l, \dots \rangle$ is a model of T . The expanded model contains interpretations for the predicate, function, and constant symbols of \mathcal{L} that do not appear in the language of the scene model.

Define the *restriction of M_S relative to T* , written as $M_S(T)$, to be the models in M_S that are compatible with T . $M_S(T)$ characterizes the total information for the scene contained within the hybrid system. The fundamental challenge is to provide mechanisms for modifying the analogical structures to reflect the contents of $M_S(T)$. In particular, we require both a *consequence operation* for determining sentences of \mathcal{L} that are logically entailed by $M_S(T)$ and an *update operation* for modifying the scene structure to reflect the derived consequences.

In general, the analogical structures may have weaker representational capabilities than is required to capture the information content of $M_S(T)$. Consider the diagram (1) and the contingent theory

$$T_0 = \{ RESIDES(Cyril, H_1), RESIDES(Paul, H_1) \}.$$

These two sources of information jointly imply that U_2 and U_3 are offices and that Paul and Cyril each own one of these offices, although it is undetermined as to who owns which one. Every model in $M_S(T_0)$ either has both $\langle U_2, Cyril \rangle$ and $\langle U_3, Paul \rangle$ or both $\langle U_2, Paul \rangle$ and $\langle U_3, Cyril \rangle$ in its *OWNS* relation. However, this information cannot be manifest in the scene structure since it is not definite about who owns which office and the scene structure does not admit disjunctive information about ownership. We shall say that an analogical structure Q *adequately represents* a set of scene models M if $M \subseteq M_Q$ and there is no other scene structure R such that $M \subseteq M_R$ and $M_R \subset M_Q$.

We will call a given consequence and update operation pair *sound* if it generates only scene structures whose models contain $M_s(T)$ and will call the pair *complete* if it produces structures that adequately represent $M_s(T)$.

3 An Integration Framework

We now present a collection of sound but incomplete integration rules for merging scene and contingent information. As will be seen, completeness cannot be attained without the addition of substantially more machinery to the basic framework.

3.1 Interface

The integration process requires two types of translation mechanisms for communicating information between the representation subsystems: *reflection* and *extraction* procedures.

Reflection procedures provide a means of inserting information into an analogical structure. For each label and analogical predicate $P(u, v)$ we require a reflection procedure $\text{INSERT}.P(u, v)$ such that $P(u, v)$ holds in all models of S after $\text{INSERT}.P(u, v)$ is invoked, for u and v in $E_s \cup E_l$.

Extraction procedures provide access to the information stored in the analogical structure for use by the sentential subsystem. To see why such access is necessary, note that the diagram (1) and the sentence

$$\forall x. \text{INHALL}(x, V) \supset \text{TYPE}(x, \text{Office}) \quad (7)$$

jointly imply that both U_2 and U_3 are offices; this conclusion cannot be deduced from (7) alone. This example illustrates the need for two-way flows of information between the sentential and analogical subsystems. In other words, assimilating sentential information into analogical structures generally requires the extraction of information from the analogical structures.

As noted above, whole-scale translation of the analogical structures into first-order sentences is infeasible. Instead, we wish to provide access to the information in the analogical structures on an *as needed* basis, whereby information is accessed as required for individual deduction steps rather than all at once. The two key types of information stored within scene structures are analogical and label relationships for scene elements, and closure information about those relationships.

For each scene predicate $P(u, v)$, we require an extraction procedure $\text{EVAL}.P(u, v)$ for evaluating ground instances in a scene S ; the procedure behaves as follows for $u, v \in E_s \cup E_l$:

$$\text{EVAL}.P(u, v) = \begin{cases} \text{true} & \text{if } M_S \models P(u, v) \\ \text{false} & \text{if } M_S \models \neg P(u, v) \\ \text{unknown} & \text{otherwise} \end{cases}$$

Let $P[x]$ represent an instance of a predicate in $\mathcal{P}_L \cup \mathcal{P}_A$ that contains the single variable x , for example, $\text{BES}(x, U_1)$. To extract information about closure relationships for a given scene S , we employ the procedures $\text{CLOSURE}^+.P[x]$ and $\text{CLOSURE}^-.P[x]$, defined as follows:

$$\text{CLOSURE}^+.P[x] = \{e \in E_l \cup E_s \mid m \models P[e] \text{ for some } m \in M_S\}$$

$$\text{CLOSURE}^-.P[x] = \{e \in E_l \cup E_s \mid M_S \models P[e]\}$$

With respect to the scene S , $\text{CLOSURE}^+.P[x]$ generates the set of scene elements that possibly satisfy $P[x]$ (called the *minimal superclosure*) while $\text{CLOSURE}^-.P[x]$ generates the set of elements that definitely satisfy $P[x]$ (the *maximal subclosure*). The procedures $\text{CLOSURE}^+.P[x]$ and $\text{CLOSURE}^-.P[x]$ give minimal upper- and maximal lower-bounds, respectively, for the *exact closure* for $P[x]$, which is the precise set of values for which $P[x]$ is satisfied. In the case where the analogical structure determines the exact closure of the predicate, that is, for every $m_1, m_2 \in M_S(t)$:

$$\forall e \in E_s \cup E_l. m_1 \models P[e] \equiv m_2 \models P[e],$$

the maximal sub- and minimal superclosures are both equal to the exact closure.

3.2 Inference Rules

The inferential component of the integration framework contains a standard proof-theory for first-order logic along with the rules of *evaluation* and *domain enumeration*. These two rules utilize information from the scene structure as provided by the extraction procedures to simplify formulas of T , eventually deducing ground instances of predicates in $\mathcal{P}_A \cup \mathcal{P}_L$ that are consequences of $M_S(T)$. These consequences are used to update the scene structures through application of the corresponding reflection procedures. In defining the two inference rules, we use the notation α_b^c to represent the expression α with all occurrences of the expression b replaced by c .

3.2.1 Evaluation The evaluation rule sanctions replacement of ground instances of a predicate in $\mathcal{P}_A \cup \mathcal{P}_L$ by either *true* or *false*, in accordance with the contents of the scene structure. In the case where the structures are incomplete and the relationship denoted by the predicate under evaluation is undetermined, the evaluation process has no effect.

Definition 3.1 (Evaluation) Let ϕ be a formula that contains an instance $R(t_1, \dots, t_k)$ of a predicate $R \in \mathcal{P}_A \cup \mathcal{P}_L$. If $\text{EVAL}.R(t_1, \dots, t_k) = \theta$ where $\theta \in \{\text{true}, \text{false}\}$ then evaluation of $R(t_1, \dots, t_k)$ in ϕ yields $\phi_{R(t_1, \dots, t_k)}^\theta$.

3.2.2 Domain Enumeration The domain enumeration rules allow the elimination of quantifiers in certain cases through the introduction of an appropriate domain of values that covers the relevant instantiations of the quantified variable. Consider the assertion

$$\exists x. \text{BES}(x, U_2) \wedge \text{OWNS}(x, \text{Cyril}) \quad (8)$$

relative to scene (1). The interpretation of this formula is that the scene element owned by Cyril is located beside U_2 . The conjunct $\text{BES}(x, U_2)$ limits the possibilities for this scene element; diagram (1) indicates that the element must be either U_1 or U_3 . As such, the formula $\text{OWNS}(U_1, \text{Cyril}) \vee \text{OWNS}(U_3, \text{Cyril})$ follows from (8). This derived formula and (8) are equivalent since $\{U_1, U_3\}$ is the exact closure for $\text{BES}(x, U_2)$.

Similarly, consider the universally quantified formula

$$\forall x. \text{INHALL}(x, V) \supset \text{TYPE}(x, \text{Office}), \quad (9)$$

which asserts that all elements in hallway V are offices. This formula can be viewed as a statement about the predicate $\text{TYPE}(x, \text{Office})$, with $\text{INHALL}(x, V)$ serving as a filter on the set of relevant instantiations of the quantified variable. According to the scene (1), the only values that satisfy $\text{INHALL}(x, V)$ are $\{U_1, U_2, U_3\}$. Thus, we can derive the conjunction $\text{TYPE}(U_1, \text{Office}) \wedge \text{TYPE}(U_2, \text{Office}) \wedge \text{TYPE}(U_3, \text{Office})$. In fact, this formula is equivalent to (9) since $\{U_1, U_2, U_3\}$ constitutes the exact closure for $\text{INHALL}(x, V)$.

We refer to the technique used above for folding in closure information as *domain enumeration* for a quantifier.² For a given quantified formula, the occurrence of a predicate instance that both contains the

²The technique of domain enumeration derives from the

variable of quantification and has its closure determined by the analogical structure is not sufficient to guarantee the applicability of domain enumeration. The formula $\exists x. \neg BES(x, U_1) \wedge TYPE(x, Closet)$ illustrates this point. In this case, the closure for $BES(x, U_1)$ is not an appropriate restriction of the terms of \mathcal{L} ; elimination of the existential quantifier from this formula using the closure would lead to unsound conclusions.

For existential quantifiers, the domain used in domain enumeration must include all bindings for which the embedded formula (e.g., $OWNS(x, Cyril) \wedge BES(x, U_2)$ in (8)) may have truth value *true* in order to guarantee that all relevant instantiations of the quantified variable are covered. For universal quantifiers, the domain should exclude values for which the embedded formula is already determined to have truth value *true*. We call a predicate instance whose exact closure satisfies these conditions *focus expressions* for the given quantified formula. In essence, a focus expression prunes from consideration those bindings of a given quantified variable that do not provide useful information. To formalize the concept of focus expressions, we introduce definitions for the *polarity* and *definiteness* of predicate instances in a formula.

Definition 3.2 (Polarity) An instance of a predicate in a formula ϕ is called positive if the instance maps to an unnegated literal in the conjunctive normal form of ϕ and is called negative otherwise.

Definition 3.3 (Definiteness) An instance of a predicate in a formula ϕ is called definite if the instance maps to a literal in a clause of length one in the conjunctive normal form of ϕ and is called indefinite otherwise.

We will combine the notions of polarity and definiteness, referring to individual instances as *negative indefinite* or *positive definite* as appropriate. The expression $INHALL(x, V)$ is a negative indefinite instance in $\forall x. INHALL(x, V) \supset TYPE(x, Office)$ and a positive definite instance in $\exists x. INHALL(x, V) \wedge TYPE(x, Closet)$.

Definition 3.4 (Focus Expression) If ψ is a quantified formula containing a predicate instance $P[z]$ then $P[z]$ is a focus expression for ψ iff either

- ψ has the form $\forall z. \alpha$ and the occurrence of $P[z]$ is negative indefinite, OR
- ψ has the form $\exists z. \alpha$ and the occurrence of $P[z]$ is positive definite.

When applying domain enumeration to a universally quantified formula $\forall z. \alpha[z]$, the embedded formula $\alpha[z]$ need not be fully retained. Instead, the simplification of $\alpha[z]$ in which the focus expression is replaced by *true* will suffice since the focus expression has truth value *true* for all terms in any of its subclosures. The focus

predicate-based generation method introduced in [8] for automatically generating attachments to evaluate quantified formulas.

expression needs to be retained for existentially quantified formulas though, since by definition the superclosure may contain terms that are not in the exact closure (and hence do not satisfy the focus expression).

The domain enumeration rule is formally defined as follows.

Definition 3.5 (Domain Enumeration) If ψ is a quantified expression, either $\exists z. \alpha$ or $\forall z. \alpha$, containing a focus expression $\Phi[z]$ with maximal subclosure D^- and minimal superclosure D^+ then domain enumeration for ψ and $\Phi[z]$ yields:

$$\bigvee_{d \in D^+} (\alpha_z^d) \quad \text{if } \psi \text{ is } \exists z. \alpha$$

$$\bigwedge_{d \in D^-} (\alpha_z^d)_{\Phi[d]}^{true} \quad \text{if } \psi \text{ is } \forall z. \alpha$$

3.3 Example

We illustrate the workings of our integration rules by applying them to the scenario presented in Section 2.2 for the scene structure of (1). Consider first the fact that Paul's office is in hall H_1 , given by the formula $RESIDES(Paul, V)$. Rewriting using definition (4) yields:

$$\exists u. INHALL(u, V) \wedge TYPE(u, Office) \wedge OWNS(Paul, u) \quad (10)$$

The predicate $INHALL(u, V)$ is a focus expression in (10) and its exact closure in scene (1) is $\{U_1, U_2, U_3\}$. Domain enumeration using this focus expression yields

$$\bigvee_{d \in \{U_1, U_2, U_3\}} TYPE(d, Office) \wedge OWNS(Paul, d) \quad (11)$$

Scene (1) contains the information that U_1 and U_3 are offices, thus $TYPE(U_1, Office)$ and $TYPE(U_3, Office)$ in (11) can be replaced by *true* using the evaluation rule. In addition, since the diagram indicates that Ralph owns U_1 , evaluation can be used to rewrite $OWNS(Paul, U_1)$ to *false*. These evaluations combined with tautological simplification produces

$$TYPE(U_2, Office) \wedge OWNS(Paul, U_2) \vee OWNS(Paul, U_3) \quad (12)$$

Similarly, from the formula $RESIDES(Cyril, H_1)$ we obtain the disjunction

$$TYPE(U_2, Office) \wedge OWNS(Cyril, U_2) \vee OWNS(Cyril, U_3) \quad (13)$$

Expansion of the contingent fact $\neg NBR(Ralph, Paul)$ using definition (3) gives

$$\forall x, y. \neg ((TYPE(x, Office) \wedge TYPE(y, Office) \wedge OWNS(Ralph, x) \wedge OWNS(Paul, y) \wedge BES(x, y)))$$

This formula contains the focus expression $OWNS(Ralph, x)$ whose exact closure in scene (1) is $\{U_1\}$; domain enumeration yields:

$$\forall y. \neg (TYPE(U_1, Office) \wedge TYPE(y, Office) \wedge OWNS(Paul, y) \wedge BES(U_1, y))$$

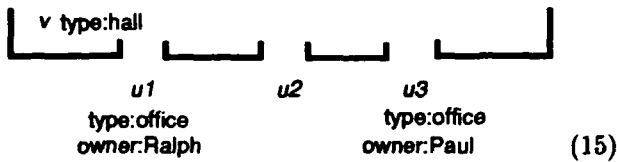
The expression $TYPE(U_1, Office)$ evaluates to *true* in the diagram; thus, we obtain

$$\forall y. \neg TYPE(y, Office) \vee \neg OWNS(Paul, y) \vee \neg BES(U_1, y). \quad (14)$$

$BES(U_1, y)$ is a focus expression in (14) and its exact closure is $\{U_2\}$; domain enumeration for $BES(U_1, y)$ yields

$$\neg TYPE(U_2, Office) \vee \neg OWNS(Paul, U_2).$$

This formula along with (12) jointly entail the conjunction $TYPE(U_3, Office) \wedge OWNS(Paul, U_3)$. The reflection operators can be applied to the conjuncts of this formula to create the scene structure:



Evaluation can be applied using this diagram to rewrite $OWNS(Cyril, U_3)$ in formula (13) to *false* since the diagram indicates that the owner of U_3 is Paul. The result is the formula $TYPE(U_2, Office) \wedge OWNS(Cyril, U_2)$. Note that this last deduction could not be made from (13) and the sentence $OWNS(Paul, U_3)$ alone; we need the further information that ownership is unique. The uniqueness property is embedded in the scene structure and is used implicitly by the extraction procedures. Finally, the contents of this last formula can be reflected to produce the final diagram (6).

3.4 Properties of the Framework

The integration framework satisfies the following properties.

Proposition 3.6 (Soundness) *Evaluation and domain enumeration are sound.*

Proposition 3.7 (Equivalence) *Domain enumeration using exact closures is an equivalence-preserving inference rule.*

The integration rules are not complete. It is shown in [8] that a hybrid system in which information is transferred into a sentential subsystem on an as-needed basis requires both closure information for predicate instances with arbitrary numbers of free variables and the integration power of *theory resolution* [9] to guarantee completeness. A framework of the type presented here could be extended for a given application to ensure completeness, although such extensions will not always be practical because of the resultant complexity of the interface between subsystems.

We note that the integration rules proposed here can all be implemented using attachment technology [8].

4 Conclusion

We have presented a formal hybrid framework for integrating domain information expressed sententially into the perceptual interpretation process and shown how

the framework extends the capabilities of traditional perception systems.

The given framework provides greater functionality than previous formalisms from the hybrid reasoning community [4] through its capacity to reflect derived information back into the analogical structures. We are currently developing more general inference rules that manipulate the scene structures directly in order to reason hypothetically. Such inference rules provide a partial solution to the incompleteness impasse described above. This research can be viewed as a computational realization of Johnson-Laird's mental models [6] and is similar in spirit to work in the Hyperproof [1] and WHISPER [5] systems, although tailored to the application of perceptual interpretation.

References

- [1] J. Barwise and J. Etchemendy. Visual information and valid reasoning. In W. Zimmerman, editor, *Visualization in Mathematics*. Mathematical Association of America, Washington, DC, 1990.
- [2] A. Bobick and R. Bolles. An evolutionary approach to constructing object descriptions. In *Proceedings of the Fifth International Symposium on Robotics Research*, Tokyo, Japan, 1989.
- [3] C. Chang and H. J. Keisler. *Model Theory*. Elsevier Press, New York, 1977.
- [4] A. M. Frisch and R. B. Scherl. A bibliography on hybrid reasoning. *AI Magazine*, 11(5), 1991.
- [5] B. V. Funt. Problem-solving with diagrammatic representations. *Artificial Intelligence*, 13, 1980.
- [6] P. N. Johnson-Laird. Mental models in cognitive science. *Cognitive Science*, 4:71-115, 1980.
- [7] A. Mackworth and R. Reiter. A logical framework for depiction and image interpretation. *Artificial Intelligence*, 41, 1989.
- [8] K. L. Myers. *Universal Attachment: An Integration Method for Logic Hybrids*. PhD thesis, Stanford University, 1991.
- [9] M. E. Stickel. Automated deduction by theory resolution. *Journal of Automated Reasoning*, 1(4), 1985.

Using Default and Causal Reasoning in Diagnosis

To appear in: *Annals of Mathematics and Artificial Intelligence*, 1993

Kurt Konolige
Artificial Intelligence Center
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
konolige@ai.sri.com

May 7, 1993

Abstract

We present a theory of default reasoning that is specifically targeted to causal domains. These domains encompass a wide variety of current applications of default reasoning, but here we concentrate on model-based diagnosis. The theory is unique in that it integrates a formal notion of causality with nonmonotonic reasoning techniques based on default logic and abduction. The main structure of the theory is a default causal net (DCN) representing the causal connections among propositions in the domain. The causal net provides a framework for the two nonmonotonic reasoning techniques of assuming defaults and generating explanations for observations, allowing them to be combined in a principled way.

1 Introduction

Knowledge of causation is an important part of commonsense reasoning. We use cause-and-effect analysis to understand everything from why we caught the flu to how to make a video recorder save our favorite TV show. If causation is so ubiquitous in reasoning about and affecting everyday events, it might also be a useful concept to employ in a formal theory of diagnosis. Surprisingly, the best-known such theory, model-based diagnosis [Reiter, 1987], does not. We argue in this paper that importing a formal notion of causation into model-based diagnosis leads to a better theory, solving some significant representational and inference problems.

What benefits can an explicit encoding of causation bring to diagnostic theories? There are at least three possible areas:

- Problem structuring
- Explanations
- Computation

The first, problem structure, is the most important, and underlies the other two. It is clear that in everyday reasoning we use the concept of cause and effect to structure our interpretations of the observations we make, to understand how events occur and how we can affect them. This representational issue is the main focus of the paper, and just below we present an example motivating our viewpoint.

The second item, explanations, is important whenever a diagnostic system must communicate its results to an end user. In answering questions about how a conclusion was reached, it is not acceptable for a system to state:

X is 13 and Y was 12 and the system equation predicts that Z will be 18.

This kind of "explanation" will not be helpful: it does not give a user insight into the domain in terms that he is familiar with, i.e., causal relations.

Finally, there are computational issues. By giving a structure to the domain, one that usually has a strong acyclic bias, causal relations can focus the computational task. Some examples of the benefits that can result are in the theory of Bayes nets [Pearl, 1988] and in using causal approximations to

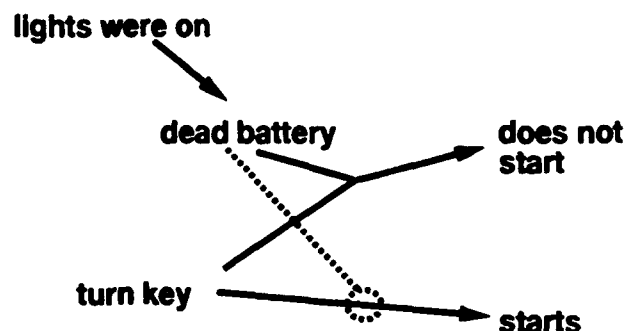


Figure 1: Starting the car

physical theories [Nayak, 1992a; Nayak, 1992b]. Although we give some computational methods at the end of this paper, these are mostly to touch base with previous work in model-based diagnosis, and we have not yet explored the computational ramifications of the theory.

To return to structural issues: it is important to understand that the utility of the concept of causation depends to a large extent on our ability to use defaults. Since the information available to us about any given situation is limited, we often must make informed guesses about the situation in order to proceed with any causal inferences. Since this talk of causation and defaults is very abstract, it will help to illustrate some of the issues involved by considering the mundane example in Figure 1. The solid arrows represent "normal" causal connections among the propositions. Turning the key will normally cause the car to start; if the lights were on overnight, there normally will be a dead battery. A dead battery means that the car will not start. There are also other kinds of information present: a dead battery blocks the causal relation between turning the key and starting the car. This information is represented by the dashed line.

Now suppose we know that the lights were on overnight, and we turn the key. What conclusions should we draw? On the one hand, we can argue that the lights were on, so the battery should be dead, and so turning the key will not start the car. This is the natural conclusion to draw; but there is another one we might argue for. Suppose we start by assuming that turning the key actually will start the car; then it can't be the case that there is a dead battery, and so perhaps leaving the lights on did not affect the battery

in the normal way. Both these arguments violate one normal condition: in the first argument, the condition that turning the key normally starts the car; in the second, that leaving the lights on drains the battery.

Intuitively, we accept the first argument because although the normal condition for starting the car is violated, there is an explanation or excuse for the violation: the battery is dead. To show this, we have drawn a dotted line in Figure 1 connecting the proposition dead battery to the causal relation between turning the key and starting the car. But there is no such excuse for concluding that the lights being on did not drain the battery. We have to invent a plausible account of how this might happen, which makes it a less persuasive argument than its competitor. In the absence of additional information, we conclude that the car will not start.

Suppose we learn that, after turning the key, the car did indeed start. Now we can no longer accept the first argument, because it leads to a conclusion we know to be false. The only other explanation of what occurred is the second argument: something must have prevented the lights being on from draining the battery.

This example illustrates some key principles of reasoning in causal domains.

- In domains where we have incomplete information, causal reasons are subject to default assumptions for their application.
- Defaults that lead to conflicting conclusions occur frequently, and the correct default can often be inferred from causal precedence among the defaults.
- Explanations for observations are generated by assuming various causal hypotheses that could lead to the observations. The most natural explanations are those that have the fewest unexplained violations of defaults.

In the sequel we present a theory of causal and default reasoning that is based on these three principles. It is important to note that the purpose is *not* to develop a theory of causation itself by reducing it to other, more primitive concepts. This is the goal of some philosophical theories of causation, e.g., Suppes [Suppes, 1970] defines causation in terms of conditional probabilities of events, or Lewis [Lewis, 1973] in terms of counterfactual statements about

possible worlds. Rather, we assume causation is a primitive relation among events, and use it to structure arguments about what defaults should apply in a given situation, and what conclusions we should accept. We call any theory that unifies causal and default reasoning a *causal default theory*. There have been other proposals to use causation in formal theories; perhaps the closest to our approach is the use of causation in Bayes nets [Pearl, 1988]. Our goal is similar: to use causation as a structuring concept to guide the application of a formalism that is, in effect, too general. Large probability distributions are difficult to define in particular domains, and causation provides an abstract view that structures the probability space with independence assumptions. Analogously, we use causation to structure logical theories of defaults, furnishing both a guide to the application of default reasoning, and a formal system that functions at a useful level of abstraction.

There are two tasks that a causal default theory should address: predication and explanation. Prediction is the process of deriving the course of events from initial conditions. Prediction is useful in many ways, for example, in planning one's actions. What happens if I don't pay my telephone bill on time? Knowing the consequences of this action can help decide whether to perform it or not. Another way prediction is used is to set up expectations in testing. An electronics engineer may apply an input to a circuit, expecting it to generate a certain output if it is working correctly.

The second task is explanation: from observed effects, infer what could have caused that effect. Typical here are applications such as plan recognition and diagnosis of complex systems. In plan recognition, one tries to infer the intentions of someone through observation of her actions: *Why did the train conductor ask if I had a passport?* Understanding the relation of actions to intentions is important in any cooperative task, and especially in communication [Cohen *et al.*, 1990]. Diagnosis is a similar kind of task, except that one is trying to figure out possible explanations for a system not behaving as expected: *Why does the copier always jam when I put in transparency paper?* Finding the answer to this question can help in fixing the problem.

Prediction and explanation are related. An explanation for an observation is a hypothesis H that, if true, would predict the observation. For a causal explanation, the connection must be stronger: H must predict the observation as a causal consequence.

Given the importance of the concept of causation, it is perhaps surprising that there is no explicit mention of causation in formal model-based diagnosis (MBD) theories (e.g., [Reiter, 1987]). In MBD, the normal functioning of a system is represented by a first-order theory SD. The interesting part of the original version of MBD is that it is not necessary to state how the system will fail. Given an observation of nonnormal behavior, a diagnosis is obtained by deciding which components would, if working correctly, contradict the observations, and hence must be broken. Later versions of MBD have added more information about failure modes [de Kleer and Williams, 1989; de Kleer *et al.*, 1990; Struss and Dressler, 1989].

From a causal point of view, the system description SD is a set of constraints between the states of components of the system and its input-output behavior, but it does not necessarily represent a causal relation. For example, given the observation that the output of an inverter circuit is logical one, MBD would predict that the circuit is functioning normally and the input is logical zero; yet the output does not cause the input. As we argue in Section 2, the lack of an explicit causal relation can be a drawback for current theories of MBD.

There are some diagnostic theories that contain an explicit causal relation, mostly connected with medical domains. Many of these are variants of the set-covering model of Reggia *et al.* [Reggia *et al.*, 1985], and often involve a probabilistic component. However, these approaches generally do not represent the normal function of a system, and are limited to expressing causation by means of a simple relation between events, rather than using a more expressive logical language, as we do here.

In this paper we present a theory that integrates causal and default reasoning within a first-order framework. Both the normal function of a system, and full or partial information about its fault modes can be represented. The main structure of the theory is a default causal net (DCN) representing the causal connections among propositions in the domain. Default causal nets, we claim, offer significant representational advantages over current formal model-based diagnosis theories.

- DCNs distinguish between the strong explanation of the cause of an observation versus the weaker explanation of an excuse for the consistency of the observation.
- Partial fault models are allowed; information about fault modes can

lead to stronger explanations, but complete information is not required.

- Preferences among explanations based on causal relations in DCNs can yield better diagnoses than current model-based theories.
- Because it is based on abductive reasoning, DCNs admit causal influences that are neither normal or abnormal, but neutral.

Some of these advantages accrue because DCNs use an abductive approach to explanation in diagnosis; others, especially the third, are a result of incorporating an explicit causal relation.

The theory of DCNs is introduced in Section 3, and their main properties explored, including the relation to MBD. In Section 4 we examine proof methods for DCNs that are similar to the familiar methods for MBD, including candidate generation and an ATMS implementation (all of the examples in this paper have been solved automatically by this implementation, although we have not tried to scale up to larger problems). Finally, we compare DCNs to more recent abductive methods in MBD, and to other formal approaches to causation that have appeared in the AI literature, and point out some of the difficulties and extensions of our approach.

2 Model-based diagnosis

Here we discuss some of the limitations of MBD that could be improved with the addition of a causality relation. We use the definitions of Reiter [Reiter, 1987]. In Reiter's theory, a system is a tuple $(SD, CMPS)$, where SD is a first-order theory describing the system and $CMPS$ is a set of component names. The distinguished predicate ab is used to describe a malfunctioning component.

Definition 1 *A diagnosis for observations O relative to a system $(SD, CMPS)$ is a minimal set of components $\Delta \subseteq CMPS$ such that*

$$SD \cup O \cup \{ab(c) \mid c \in \Delta\} \cup \{\neg ab(c) \mid c \in CMPS - \Delta\}$$

is consistent.

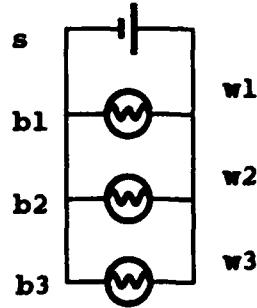


Figure 2: Three Bulbs Example

2.1 Excuses vs. explanations

Reiter's definition identifies a diagnosis as a type of excuse. Consider the example of Figure 2 (adapted from [Struss and Dressler, 1989]). There are three bulbs in parallel with a source. Let us suppose for simplicity that the wires always behave correctly. The system description is:

$$\neg ab(s) \wedge \neg ab(b_i) \supset on(b_i), \quad i = 1, 2, 3 \quad (1)$$

If we observe that b_3 is off, there are two diagnoses, $\{s\}$ and $\{b_3\}$. This is intuitively correct, and we might be tempted to say that the abnormality of s causes b_3 to be off. But the formalism does not support such a view: there is no prediction from $ab(s)$ to bulb b_3 being off, only the weaker excuse that if the source is abnormal, it is not inconsistent that b_3 is off. An explanation of why b_3 is off should involve a prediction from the hypotheses (i.e., the diagnosis).

To see the difference between explanations and excuses in another way, we add the observation that b_2 is on. It is a curious fact of MBD that the diagnoses do not change, although we might expect $\{s\}$ to be dropped. But, when the source is abnormal, the system description is consistent with *any* state of the bulbs; hence, it is an excuse for the observations, even though it does not explain why the second bulb is on (or why the third one is off).

There are several ways of getting rid of the unfortunate diagnosis. The simplest solution is to add so-called "physical impossibility axioms" [Friedrich *et al.*, 1990]. For example, in the three bulbs case, the axioms:

$$\neg(on(b_i) \wedge ab(s)), \quad i = 1, 2, 3 \quad (2)$$

are constraints that state a bulb cannot be on without a source of power. It is easily checked that $\{s\}$ is no longer a diagnosis for $\{on(b_2), \neg on(b_3)\}$.

While the physical impossibility axioms are helpful, some problems remain in identifying diagnoses with excuses. First, all of the necessary axioms must be included, and there are no guidelines for determining what they are. For complex systems, it may be difficult to determine the physical impossibility axioms, since we must enumerate all combinations of atoms that are physically impossible. Second, the diagnoses are still excuses: as in the case of $\{b_3\}$, they make no prediction of the observations from hypotheses.

Another solution is to add fault models, the solution proposed by Struss and Dressler [Struss and Dressler, 1989]. In this case, there can be a prediction from a hypothesized abnormality to the observed behavior. For the three-bulb example, appropriate fault axioms are:

$$ab(b_i) \supset \neg on(b_i), \quad i = 1, 2, 3. \quad (3)$$

With this addition we can predict the observation $\neg on(b_3)$ from the diagnosis $\{b_3\}$.

The inclusion of fault models brings MBD closer to a causal viewpoint. but there are still problems and discrepancies. As we discuss in some detail below, often we do not have full information about failure modes of a system, and it is problematic to include partial fault models into MBD. And even if full fault models are included, MBD does not generate causal explanations or predictions.

In contrast, a causal theory offers a more reasoned approach to diagnosis, simulating the intuitive process by which we arrive at the correct solution. For the bulb example, we represent that the source normally causes b_1 , b_2 , and b_3 to be on. In seeking to explain the observations $on(b_2)$ and $\neg on(b_3)$, the most normal explanation is that the source and b_2 are working correctly (thereby explaining causally why b_2 is on), and that b_3 is broken, thereby excusing the normal causation of b_3 being on. In this example, there is no need for fault models or physical impossibility axioms: the process of finding causal explanations and excuses generates the correct diagnosis.

2.2 Fault models, relevance, and neutral causes

As we have seen, fault models are necessary if observations are to be predicted or explained, rather than excused. But there are some well-known problems

with using fault models in MBD. The first is that partial fault models are generally useless, as was pointed out in [de Kleer and Williams, 1989]. A partial fault model is one in which we have knowledge of some of the failure modes of a component, but not all of them. Returning to the AND gate example, suppose that ignoring the x input is only one way in which the gate can fail; call this failure mode $f(a)$. Then the system description becomes:

$$\begin{aligned} f(a) \supset o &= y \\ f(a) \supset ab(a) \\ \neg ab(a) \supset o &= x \times y \end{aligned} \tag{4}$$

Again observing $o = 1$ and $x = 0$, there is a diagnosis $\{a\}$. But since $f(a)$ is only one failure mode, it is not implied by $ab(a)$, and no further predictions can be made. Partial fault models are never used for inference, because there is always the possibility of the component failing in some other, unknown way. De Kleer and Williams [de Kleer and Williams, 1989] propose using *behavioral modes* in place of abnormalities, that is, a component will have a set of failure modes that are known and described with axioms (e.g., the AND gate short in (4)), and perhaps an unknown mode with no description. Diagnoses are combinations of normality assumptions and fault modes.

Whenever fault models are used, minimal diagnoses no longer cover the set of all diagnoses for a system. This problem was noted and discussed in [de Kleer *et al.*, 1990], and we review it briefly here. As long as no fault models are present, all occurrences of ab in the system description are positive (i.e., in axioms such as $\neg ab(a) \supset o = x \times y$). In this case, a diagnosis Δ , which is a minimal set of abnormal components, actually stands for or covers a whole set of diagnoses, namely any superset of Δ . Δ is the relevant core of all of these nonminimal diagnoses.

Adding fault models changes this picture, since supersets of a diagnosis are no longer guaranteed to be diagnoses. Instead, in [de Kleer *et al.*, 1990] the rather cumbersome construct of kernel diagnosis is substituted, and the compact covering representation is lost. More importantly, when there are axioms relating abnormalities, as in Equation (6), MBD does not distinguish the relevant failing components using either the original definition of diagnosis or that of kernel diagnosis. By relevant components, we mean those that actually predict the observations; it may happen that only a subset of the

components mentioned in a diagnosis are relevant in this way, while the others are "side effects," and have no bearing on the observations. This problem is highlighted in [Konolige, 1992], where it is shown to be inherent to the use of excuses rather than explanations.

Another related problem is the representation of components that do not have abnormal modes, i.e., a switch can be open or closed, but neither of its states is "abnormal." We call these *neutral* causes, since their presence or absence is unbiased in the system description. For example, the AND gate of Figure 3 has two neutral causes, the input lines of the gate. The redefinition of MBD in terms of kernel diagnoses in [de Kleer *et al.*, 1990] moves in the direction of allowing neutral causes, since the normal and failing modes of a component are given equal status in diagnoses. However, there is still no clear distinction between causes that have a normal or nonfailing mode, and neutral causes, which do not have a preferred mode. Finally, if the system description is complete, then neutral causes do not have to be represented at all, since their status can be predicted from the state of the components.

The problems of partial fault models, relevance, and neutral causes can be addressed by using an abductive method in place of consistency, so that the observations must follow from the system description and a hypothesized set of abnormalities. There are several approaches that differ in their details [Console and Torasso, 1991; Poole, 1989; Poole, 1993; Dressler and Struss, 1992; Besnard and Cordier, 1993]. These abductive methods are similar to the DCN framework (which also uses abduction); however, like consistency-based MBD, they have no explicit concept of causation. As we show in the next subsection, adding a causal relation leads to a better concept of faults and diagnosis, especially in terms of preferences among predictions and explanations. A detailed comparison of particular abductive methods and the DCN approach is in Section 5.

2.3 Causes vs. correlations

In MBD, it is tempting to differentiate the stronger concept of explanations from excuses by taking them to be diagnoses that imply the observations, as can be done when fault models are present. But although a fault model might *predict* a proposition, it does not necessarily give a causal explanation for the proposition. To see this, consider the simple AND circuit diagrammed in Figure 3. Assume that the only possible fault causes the input on line x

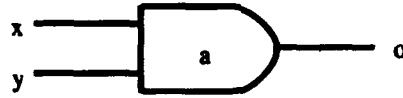


Figure 3: AND Circuit

to be ignored, so that the output and y are equal. The normal and fault axioms are:

$$\begin{aligned} ab(a) \supset o &= y \\ \neg ab(a) \supset o &= x \times y \end{aligned} \quad (5)$$

Now suppose we observe the output at 1 and x at 0. There is only one diagnosis, $\{a\}$, and from this we can predict that $y = 1$. So y is correlated with o , and we can use this information to make predictions; but it would be incorrect to say that this is a *causal* explanation for y 's value, since we know that y is an input to the device.

One might be tempted to say "so what?" at this point: is it critical to differentiate causal explanations from mere correlations in a technical theory? Our answer is yes, given the importance of causal precedence that we pointed out in the Introduction. When default assumptions are part of diagnostic reasoning, causal relations play a key role in ranking competing predictions and hence diagnoses.

Example 1 *This is strictly a prediction problem, the car example from Figure 1. In MBD, the system description is:*

$$\begin{aligned} k \wedge \neg ab(1) &\supset s \\ k \wedge d &\supset \neg s \\ l \wedge \neg ab(2) &\supset d \\ d &\supset ab(1) \end{aligned} \quad (6)$$

where k is turning the key, l is the lights were on, s is the car starting, and d is the battery being dead. Assume that l and k are known to be true, and consider them as initial conditions rather than observables to be explained. What does MDB predict in this case? Since it is impossible for both $ab(1)$ and $ab(2)$ to be false, there are two diagnoses: $\{1\}$ and $\{2\}$. Only the first of these corresponds to a causal prediction.

Example 2 This is a diagnosis example from the circuit domain. A circuit containing two inverters and an AND gate is diagrammed in Figure 4. The two inverters are coupled so that whenever b 's output is at 1, it causes a to become shorted. The system description is:

$$\begin{aligned}
 \neg ab(a) \supset x' &= 1 - x \\
 \neg ab(b) \supset y' &= 1 - y \\
 ab(a) \supset x &= x' \\
 ab(b) \supset y &= y' \\
 y' = 1 \supset ab(a) \\
 \neg ab(c) \supset z &= x' \times y'
 \end{aligned} \tag{7}$$

Suppose that initially it is known that $x = 0$ and $y = 0$, and it is observed that $z = 0$. The most likely diagnosis is that b is functioning normally, and this caused $y' = 1$ and hence the fault in a . It could be that b is faulty, but given low initial likelihood of a fault on its own, the preferred explanation is that b is functioning normally.

There are two diagnoses, $\{a\}$ and $\{b\}$; the axiom $y' = 1 \supset ab(a)$ does not differentiate them. This is because it states a correlation between y' and $ab(a)$, rather than a causal relation. Rewriting it using the axiom $\neg ab(b) \supset y' = 1 - y$, we get:

$$y = 0 \wedge \neg ab(b) \supset ab(a),$$

which in turn can be written as

$$y = 0 \wedge \neg ab(a) \supset ab(b).$$

The symmetry between $ab(a)$ and $ab(b)$ is clearly evident.

In both these examples, causal precedence arises when the normal function of one component causally entails an abnormality in another one. Writing this entailment using material implication, as is natural for the system description in MBD, does not produce the intended effect, because the implication is invertible:

$$\neg ab(a) \supset ab(b) \Rightarrow \neg ab(b) \supset ab(a).$$

A causal relation between these two would not be invertible.

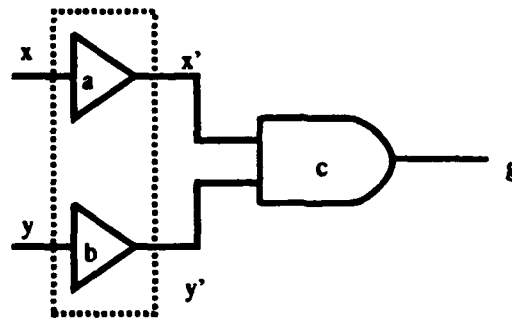


Figure 4: Two inverters and an AND gate

One extension to MBD permits preferences among the defaults [Junker, 1993; Dressler and Struss, 1992]. In this case, the correct diagnosis could be generated by assigning $\neg ab(b)$ a higher priority than $\neg ab(a)$. However, this priority does not represent a causal connection, and is inappropriate in other situations, e.g., if the input to b is 1. To have the effect of causal precedence, a diagnostic theory must explicitly code the causal relation between propositions.

2.4 Physics and causation

Physical laws, such as conservation of energy or Newton's law relating force, mass and acceleration, are written in terms of equations among state variables. Causation is not a fundamental concept, although it is often used to explain or understand physical systems. For example, Newton's law $F = ma$ says nothing about whether force causes an acceleration. Yet this is typically the way we design systems, since it is usually easy to change the force that is applied (by changing the power to a motor, for instance). The force becomes an *exogenous* variable in the equation, one that can be independently controlled, and so the direction of causation is taken to be from force to acceleration. There is nothing fundamental in such a choice; if there were a device that controlled acceleration, acceleration might be considered the exogenous variable.

In designed systems, the choice of exogenous variables, and hence the causal structure, is part of the design. Digital circuit devices are a good example. An AND device is designed so that by changing the inputs, the

output is made to vary. We can thus say that the AND gate, in normal operation, causes its output to be a boolean function of the input. An internal fault can change the function: for example, if there is an internal short, the output may take on the value of one of the inputs. In this case, the causal view of the device is unchanged, even though its functioning is abnormal.

One objection to the use of causal models in diagnosis is that they may not satisfy a modularity property, often stated as the slogan "no function in structure" [Davis and Hamscher, 1988]. In the practice of MDB, a system is composed of subsystems that have their own descriptions; the system description is generated by taking the union of all subsystem descriptions, and adding some axioms describing the connections between them. There is no reason this same modularity technique could not be used with causal models. Every subsystem would contain its own causal and correlational axioms, normal conditions and primitive causes. In addition, there would be information about the I/O behavior of the subsystem: which variables are subject to be changed by outside events (the input variables) and which are determined by other variables of the subsystem (the output variables). Causal subsystems would be hooked together in the same way as modules in the MBD approach, with the additional constraint that the I/O behavior of the subsystems must be respected.

However, there may be faults in which the causal view is undermined because the design parameters are exceeded. For example, if the output of a gate is loaded by too much fanout, then it may become stuck at zero, violating the intended behavior of the device. In this case, it is not an internal abnormality of the AND gate itself, but a violation of the design conditions under which the output is a function of the input. There are two ways to cope with this situation. One is to explicitly encode design violations using another set of causal rules. The drawback here is that these rules violate the object-centered nature of modeling devices because they arise from an interaction among the components.

Another method would be to model the device at a finer level of detail. Instead of using gates as components, the gates themselves would be modeled as collections of smaller components (resistors, transistors, etc.) together with the voltage and current relationships among them. This representation is more "physics-like," so that Kirchoff's and Ohm's laws apply, giving a complete equilibrium description of the device.

Like causal descriptions, physical descriptions also have contexts under

which they are appropriate. Actual resistors are not perfect, and there may be noise or dynamic effects in the circuit, violating the conditions for application of Kirchoff's and Ohm's laws. A finer level of description is possible at the atomic level, taking into account the material composition of the device and the flow of electrons. And so on, down to the best current physical theories at the quantum level. The point is that physical descriptions are abstractions from reality, ignoring some details and valid only within a limited context. Note that by adding more detail, more types of faults can be analyzed using physical laws; concomitantly, the complexity of description increases and the diagnosis problem becomes harder. A causal view might still be appropriate when the conditions of physical analysis are violated, and going to a finer level of detail would be too costly.

Generally speaking, causal descriptions are useful at higher levels of abstraction, especially when the intended functional dependence is known from design. This is not to say that causal models are incompatible with physical models. In the DCN theory, they coexist, with the physical description modeled as a set of statements about correlations. At higher abstraction levels, causal models are useful in directing and focusing the search for diagnoses, and in producing understandable explanations. And when full physical models are not available, causal models may be the best way to formalize the behavior of a device.

3 Default causal nets

Default causal nets (DCNs) are a formal structure that encode the concepts of causation, correlation, and defaults. They consist of a causal theory R , a definitional theory D , and a correlation or integrity theory I . In addition there are distinguished sets of propositions C (the primitive causes) and N (the normal conditions). The term "net" is used in analogy with Bayesian nets, because the main structuring concept is the causal relation embodied in R .

Definition 2 (Default Causal Net)

A default causal net is a tuple $\langle R, D, I, C, N \rangle$, where R is a Horn theory, D and I are first-order theories, and C and N are disjoint sets of atoms.

3.1 Causation

Formally, we understand causation to be a primitive relation among propositions. By "primitive" we mean that, as far as DCNs are concerned, the causation relation is part of the parameterization of the net, and is not derived from any other concepts. This is unlike the approach of Shoham [Shoham, 1987], for example, in which a theory of causation is developed by reducing it to other concepts. Our approach leaves unanswered questions about how to identify causation in a given domain, the relation of causation to time, and various other difficulties about the nature and properties of causation.

In Section 6 we return briefly to these questions to point out some difficulties of the theory; here we give the most basic (and hopefully noncontroversial) properties. These properties suffice to develop the main features of DCNs; further investigations will have to confront some of the harder problems.

To represent the causal relation, we use a definite clause theory R over a first-order language \mathcal{L} . This theory consists of a set of implications

$$a_1 \cdots a_n \supset b.$$

where each of a_i and b is a ground atom of \mathcal{L} . If A is a set of propositions, then we say that an atom b is caused by A if there is a proof of b from A in R ; we write this as $A \vdash_R b$. A is a minimal cause for b if there is no other cause A' for b such that $A' \subset A$.

Example 3 *A variation of the 3-bulb example is diagrammed in Figure 5. There is a switch that can be either open or closed. For each of the other components c_i , the proposition $ok(c_i)$ means that the component is working, and $ab(c_i)$ that it is broken. The theory R is:*

$$\begin{aligned} & \text{closed}, ok(s), ok(w_1), ok(b_1) \supset on(b_1) \\ & \text{closed}, ok(s), ok(w_1), ok(w_2), ok(b_2) \supset on(b_2) \\ & \text{closed}, ok(s), ok(w_1), ok(w_2), ok(w_3), ok(b_3) \supset on(b_3) \\ & \text{open} \supset off(b_1) \\ & \text{open} \supset off(b_2) \end{aligned} \tag{8}$$

We have not listed any fault models, although we could. Here is a partial

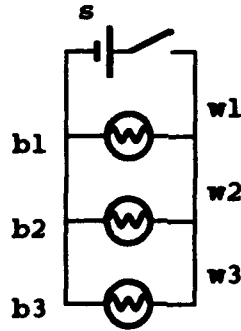


Figure 5: Three Bulbs with a Switch

fault model that we will use in some examples.

$$\begin{array}{ll}
 ab(b_1) \supset off(b_1) & ab(b_2) \supset off(b_2) \\
 ab(w_1) \supset off(b_1) & ab(w_2) \supset off(b_2)
 \end{array} \tag{9}$$

The partial fault model is also part of the relation R , since it represents causation in the abnormal functioning of the device. The primitive causes C are $\{open, closed, ab(c_i)\}$.

Note that, unlike model-based diagnosis, there can be causes other than the normal or abnormal functioning of a component. This is useful in representing neutral situations, e.g., the switch is not normally either closed or open, but can be hypothesized as either in order to explain the observations. The propositions $ok(x)$ are not listed as primitive causes; they are normal conditions, explained below.

The important part of the causal relation is that it captures the functional dependence of the domain variables. If we want to turn b_1 on, then we can close the switch and make sure that s , w_1 , and b_1 are working correctly. On the other hand, we cannot make b_1 be on as a means of causing the switch to close. Of course, if we observe b_1 to be on, then we can infer that the switch is closed; but it is not possible to *plan* to change the position of the switch by the primitive action of making the bulb be on. This illustrates the difference between a causal relation and a merely correlational one. Unlike material implication, the causal relation is asymmetric and does not contrapose: given that c causes d , it is not necessarily the case that $\neg d$ causes $\neg c$. Deduction

in a definite clause theory is one way to represent the asymmetric causal relation.

3.2 Definitions and correlations

Besides causation, there are other types of relations connecting propositions. Definitional information relates propositions that have defined relations within a domain, e.g., "a 40-watt bulb is a type of bulb" or "abnormal is the opposite of normal." Definitions can obviously interact with causation, since from "a broken 40-watt bulb caused the problem" we can infer "a broken bulb caused the problem." For our purposes, we limit definitions to information about complementary propositions. Definitional relations are represented by a first-order theory D ; for the bulbs example of Figure 5, it contains the propositions:

$$\begin{aligned} open &\equiv \neg closed \\ ab(c_i) &\equiv \neg ok(c_i) \\ on(b_i) &\equiv \neg off(b_i) \end{aligned} \tag{10}$$

If $p \vdash_D \neg q$, then we say that q is the complement of p , and write it as \bar{p} .

Information about co-occurrences is another form of non-causal information in a domain, e.g., "Whenever I clean my car it rains." Correlations can be used to make predictions, but do not contribute to causal explanations. Correlations are represented by a first-order theory I (for *integrity* theory). All causation and definition relations are also correlational. We enforce this restriction by demanding that $R \subseteq I$ and $D \subseteq I$.

Example 4 *Continuing the bulbs example, suppose we know that whenever b_1 is off and is not broken, the other bulbs must be off too. We represent this as*

$$off(b_1) \wedge ok(b_1) \supset off(b_2) \wedge off(b_3) \in I \tag{11}$$

Correlations may come from many different sources. As in the case of this example, there may be underlying but unknown causes that link together several propositions. Or we may have experiential knowledge that is the converse of causation: whenever the road is wet, it normally rained the previous night.

A proposition q is correlationally inferred from a set of propositions A if it follows logically from the correlational theory and A ; we write $A \vdash_I q$. For example, $off(b_2)$ is inferred from $A = \{ok(b_1), off(b_1)\}$ in the above example, but it is not caused by A . If A causes q , then it also infers q , since $R \subseteq I$. Note that, unlike the case with the causal relation, the material conditional can be used in for "backwards" inference, e.g., if $on(b_2)$ is true, we can infer that one of $ab(b_1)$ or $on(b_1)$ is true by using the contrapositive of Equation (11).

3.3 Normal conditions

Normal conditions are propositions that are normally assumed to hold. They generally represent either the normal functioning of a component, or a complex set of conditions, e.g., "if the key is turned and *everything is normal*, the car will start." Formally, normal conditions are a set of ground atoms N that are not primitive causes. Primitive causes are hypotheses that incur a cost to assume; normal conditions are "free" and assumed to hold by default.

Example 5 *Continuing the bulbs example, we let the set of normal conditions $N = \{ok(c_i)\}$. In this case, the normal conditions just describe the correct functioning of the components. We can define other types of normal conditions, for example to relate causation among abnormal components. Suppose that normally when b_1 is on, it causes b_2 to fail. We would write:*

$$n \wedge on(b_1) \supset ab(b_2) \quad (12)$$

as part of the causal theory R , where $n \in N$ is a new proposition reflecting a normal causal relation between b_1 and b_2 . As we will show later, such causal relations can be used to specify priorities among explanations.

Identifying normal conditions is the key to default reasoning in causal theories. We seek to explain a set of observations by hypothesizing causes that are as "normal" as possible, that is, conflict with the fewest normal conditions.

It is helpful to view the causal relation and normal conditions as a directed graph. For example, the normal functioning of the bulbs with the switch closed (Equation 8) and the failure mode just given (Equation 12) can be diagrammed as in Figure 6. The arrows show the causal connections among

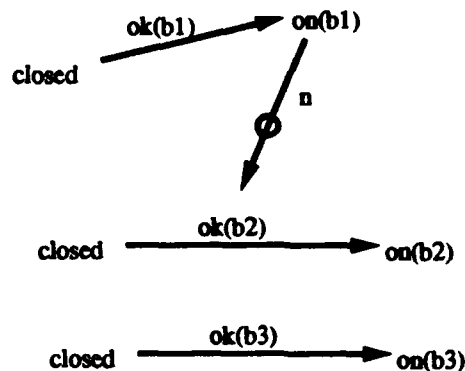


Figure 6: Causal Directionality

propositions, annotated with their normal conditions (for simplicity we have omitted some irrelevant normal conditions). The circled arrow indicates that bulb 1 being on is the cause of an abnormal condition with bulb 2. The causal directionality is clear from the diagram.

The choice of what conditions are assumed to be “normal” or part of the causal background is an important part of the information provided by the application developer. Depending on the task and the level of expertise of the developer, very different choices could be made, even in the same domain. For example, a typical driver might infer that turning the key causes the car to start, given the normal condition that the car is ok. A car mechanic might have a more detailed causal view: turning the key and having a charged battery causes the car to start, assuming the starter motor is working correctly.

3.4 Explanations

We now have all of the elements necessary to develop the inference operation of explanation within DCNs.

Definition 3 (Explanation)

An explanation for an observation set O is a set of causes and normal conditions $A \subseteq C \cup N$ such that $A \vdash_R O$ and $A \cup O \not\vdash_I \perp$.

Example 6 *To illustrate the concept of explanation, we consider the bulbs theory containing the normal causal rules (8) together with the fault model*

(9). *The fault model is necessary to provide interesting explanations of non-normal behavior. Suppose we make the observation that bulb b_1 is not lit: $\text{off}(b_1)$. There are several explanations for this proposition.*

open, ok(s), ok(b_1), ok(w_1)

closed, ok(s), ab(b_1)

etc.

There are usually many explanations for a given observation set, and we seek intuitively preferred explanations. To find these, we filter all explanations by a two-step process.

1. Normal explanation: those explanations that satisfy a maximal set of normal conditions.
2. Ideal explanation: normal explanations that have a minimal number of primitive causes.

The concept of a normal explanation is complicated by the presence of causation. An abnormal condition may be caused by the explanation; when this happens, we say that the normal condition is *exempted*. A normal explanation should either consistently include or exempt as many normal conditions as possible. Here we are using the concept of causation to structure the defaults. If a normal condition is not contained in an explanation, it counts against the explanation, *unless* the corresponding abnormal condition is exempted.

Definition 4 (Adjunct)

Let A be an explanation for observation set O . The adjunct of A is a set of normal conditions defined as follows.

- *If the complement \bar{x} of a normal condition x is in A , then x is in the adjunct.*
- *If a normal condition x is not in A , and $A \not\models_R \bar{x}$, then x is in the adjunct.*

A normal explanation for O is one whose adjunct does not strictly contain the adjunct of any other explanation for O . An ideal explanation is a normal one that is subset-minimal in the primitive causes.

Example 7 As in the previous example, consider the bulbs theory (8) together with the fault model (9). Again, if we make the observation that bulb b_1 is off, we have several candidates for normal explanations:

<i>Explanation</i>	<i>Adjunct</i>
$ok(s), open, ok(w_1), ok(b_1) \dots$	<i>none</i>
$ok(s), ab(w_1), ok(b_1) \dots$	$ok(w_1)$
$ok(s), ok(w_1), ab(b_1) \dots$	$ok(b_1)$

Of these, the minimal adjunct is the first. This is the normal and ideal explanation of $off(b_1)$: the switch is open, and all components are normal.

This example illustrates one property of normal explanations: as many normal conditions are assumed to hold as possible. The switch can be either open or closed; if we assume that it is open, then we have an explanation for b_1 being off that is consistent with the normal functioning of the circuit. Any other explanation will force us to assume that some component is functioning abnormally. So, normal explanations consist of a set of primitive causes that explain the observations, and at the same time respect our ideas about what normally occurs as much as possible.

In this example, there were no interesting causal relations between normal conditions. In the definition of adjunct, we used the principle of *causal exemption*: if an abnormal condition is caused by the hypothesized explanation, then it is exempted from consideration in finding the "most normal" explanation. The following example illustrates this point.

Example 8 Consider the same fault model as in Example 7 with an initial condition closed and the additional causal rule (12): $n \wedge on(b_1) \supset ab(b_2)$. There are several candidates for normal explanations of $\{off(b_2)\}$:

<i>Explanation</i>	<i>Adjunct</i>
$n, ok(s), ok(w_1), ok(b_1), ok(w_2) \dots$	<i>none</i>
$n, ok(s), ok(w_1), ok(b_1), ab(w_2) \dots$	$ok(w_2)$
$ok(s), ok(w_1), ok(b_1), ok(w_2), ab(b_2) \dots$	$ok(b_2), n$
<i>etc.</i>	

Of these, the first is the only normal explanation, and hence ideal. The reason it has an empty adjunct is that the normal conditions and closed cause $on(b_1)$, which in turn causes $ab(b_2)$, exempting the normal condition $ok(b_2)$. Every other explanation violates at least one normal condition without exempting it. This makes intuitive sense: if the switch is closed, we expect b_1 to be on, causing b_2 to be broken and off.

This example illustrates how directionality in the causal relation is important in producing causal preferences among explanations. Referring back to Figure (6), it is easy to see from following the causal arrows that $closed$, $ok(b_1)$ and n are a cause of $ab(b_2)$. On the other hand, $closed$ and $ok(b_2)$ are inconsistent with n and $ok(b_1)$, but they do not cause the complement of either of these normal conditions.

3.5 Excuses

One problem with causal explanations is that they always require a causal model that infers the observations. Without the partial fault model of Equation (9), for example, there are no explanations for why the bulbs are off when the switch is closed. In many cases, it may not be possible to find a causal explanation for all the observations, given a causal theory with incomplete fault models. In this situation the weaker concept of an excuse (discussed in Section 2.1) might be appropriate. By hypothesizing primitive causes, the normal state of the system can be changed so that it no longer conflicts with the observations.

Definition 5 (Excuse)

An excuse for observations O is a set of causes and normal conditions $A \subseteq C \cup N$ such that $A \cup O \not\models_I \perp$.

Excuses are like explanations, except there is no necessary causal relation to the observations. The following fact is obvious when the two definitions are compared.

Fact 1 *Every explanation of O is an excuse for O , but not necessarily the converse.*

Normal and ideal excuses can be defined in exactly the same manner as for explanations. Note that, although we do not use the causation relation

to infer the observations, we still use it to give preferences on the normal conditions present in excuses, in exactly the same manner as for explanations.

With excuses, we do not need to define fault models in order to "excuse" a set of observations. Excuses are useful precisely in those cases where we do not have enough information to make a predictive fault model; all we know is that some component is faulty, and we no longer can predict that the behavior of the system is at odds with the observations.

Example 9 Consider the simple bulbs theory from Equation (8), without the clauses for open. There is no explanation for $\text{off}(b_1)$, but there are several excuses:

<i>Excuse</i>	<i>Adjunct</i>
$\text{ok}(s), \text{ok}(w_1), \text{ok}(b_1) \dots$	<i>none</i>
$\text{open}, \text{ok}(s), \text{ok}(w_1), \text{ok}(b_1) \dots$	<i>none</i>
$\text{closed}, \text{ok}(w_1), \text{ok}(b_1) \dots$	$\text{ok}(s)$
$\text{closed}, \text{ok}(s), \text{ok}(b_1) \dots$	$\text{ok}(w_1)$
<i>etc.</i>	

The first two of these are normal excuses because they have minimal adjuncts. Of these, the first is ideal, because it does not have any assumed primitive causes. Note that it is unnecessary to assume open as a hypothesis, since it is predicted by the observations and the ideal excuse.

As we pointed out in Section 2.1, excuses are the idea behind Reiter's model-based diagnosis method [Reiter, 1987]. In fact, we can show that his concept of diagnosis is exactly the concept of an excuse with no causal knowledge. Recall that, in Reiter's theory, a system is a tuple $(SD, CMPS)$, where SD is a first-order theory describing the system and CMPS is a set of component names. In DCN terms, the system description corresponds to the correlational theory, the abnormalities are primitive causes, and their complements are normal conditions. The causal relation is empty; according to our analysis, Reiter's theory does not distinguish causation from correlation, all relationships are treated as correlations. In this case there are simplifications in the DCN: the adjunct of an excuse A is just the set of normal conditions not in A , and a normal excuse contains a maximally consistent set of normal conditions. We can show that normal excuses are exactly Reiter's diagnoses.

Fact 2 (Model-based diagnosis)

Let $(SD, CMPS)$ be a system. Construct a corresponding DCN as follows:

$$\begin{aligned}
 R &= \emptyset \\
 I &= SD \\
 D &= \{ab(c) \equiv \neg ok(c) \mid c \in CMPS\} \\
 C &= \{ab(c) \mid c \in CMPS\} \\
 N &= \{ok(c) \mid c \in CMPS\}
 \end{aligned}$$

Then Δ is a diagnosis of O with respect to $(SD, CMPS)$ if and only if $\{ok(c) \mid c \in CMPS - \Delta\}$ is a normal excuse for O in the corresponding DCN.

Proof. Assume that Δ is a diagnosis of O . Then $\{\neg ab(c) \mid c \in CMPS - \Delta\}$ is a maximal set of normal predicates consistent with SD . Therefore, from the definitional theory, $\{ok(c) \mid c \in CMPS - \Delta\}$ is a maximal set of normal conditions consistent with I , and since the causal relation is empty, this is a normal excuse.

In the other direction, assume X is a set of components such that $\{ok(c) \mid c \in X\}$ is a normal excuse for O . Again from the definitional theory, $\{\neg ab(c) \mid c \in X\}$ is a maximal set of normal predicates consistent with SD , so that $CMPS - X$ is a minimal set of abnormality predicates consistent with SD .

Because the model-based theory does not have a causation relation, causal exemption and preferences are not possible. Let us reconsider Example 8, in which the switch is closed and the observation is $off(b_2)$. Now assume that the causal rules (8) and (12) are purely correlational, and the causal theory is empty. Then we have the following excuses for $off(b_2)$:

Excuse	Adjunct
$n, ok(s), ok(w_1), ok(b_1), ok(w_2) \dots$	$ok(b_2)$
$ok(s), ok(w_1), ok(b_1), ok(w_2), ok(b_2) \dots$	n
$n, ok(s), ab(w_1), ok(b_1), ok(w_2), ok(b_2) \dots$	$ok(w_1)$
etc.	

These are all normal excuses. Either bulb b_2 is broken, or the normal connection between b_1 and b_2 doesn't hold, or wire w_1 is broken. The correct causal solution is the first one, but it cannot be distinguished from the other excuses without causal precedence.

A quick look at the relation of DCNs to MBD reveals that primitive causes and normal conditions are linked in the way that we criticized in Section 2.2.

To the extent that a DCN relies on excuses in inferring diagnoses, it is subject to many of the same criticisms made against MBD. Still, these criticisms are blunted somewhat because of the greater expressiveness of DCNs. Causal preferences can still operate even for excuses, if the causal relation is not empty; and we can make distinctions between normal conditions and primitive causes. Furthermore, it is often possible to mix explanatory and excusing components in the same diagnosis, as we show in the next section.

3.6 Lenient explanations

Excuses are weaker than explanations, and we seek explanations whenever possible, as being more informative. While there may be no explanation that covers every member of an observation set O , it may be possible to find an explanation for a subset of O , while excusing the rest.

Definition 6 (Lenient explanation)

Let $O' \subseteq O$ be a maximal subset of O for which an explanation exists. Then a lenient explanation for O is a set of causes and normal conditions $A \subseteq C \cup N$ such that A is an explanation for O' and an excuse for $O - O'$.

Obviously, if O has a causal explanation, all lenient explanations are explanations. As with ordinary explanations and excuses, we can define the concept of a normal lenient explanation and ideal lenient explanation.

Example 10 *This is the original bulb example cited in the Introduction; its causal relation is given by Equation (8), with the proposition closed assumed as an initial condition.*

Suppose we observe that bulbs one and two are off, and bulb three is on ($O = \{\text{off}(b_1), \text{off}(b_2), \text{on}(b_3)\}$). There is no explanation for $\text{off}(b_1)$ and $\text{off}(b_2)$, but there is for $\text{on}(b_3)$. So any lenient explanation of O will include $\{\text{ok}(s), \text{ok}(w_1), \text{ok}(w_2), \text{ok}(w_3), \text{ok}(b_3)\}$. In fact this is the only explanation,

since adding either $ok(b_1)$ or $ok(b_2)$ will contradict the observations. It is lenient, normal, and ideal.

Looking at the simple excuses for O , we get the following normal ones:

1. $ok(w_1), ok(w_2), ok(w_3), ok(b_1), ok(b_2), ok(b_3)$
2. $ok(s), ok(w_2), ok(w_3), ok(b_1), ok(b_2), ok(b_3)$
3. $ok(s), ok(w_1), ok(w_3), ok(b_2), ok(b_3)$
4. $ok(s), ok(w_1), ok(w_2), ok(w_3), ok(b_3)$

Each of these corresponds to a maximal set of normal conditions that does not infer $on(b_1)$ or $on(b_2)$ or $off(b_3)$; but only the last one corresponds to the correct causal explanation.

The lenient ideal explanations (LIEs) of an observation set are the ones we usually want. However, this relies on the causal model being complete for the observations. Suppose g is a member of O , and there is a causal explanation for g . If g is explained in every lenient explanation, it may conflict with reasonable excuses for other members of O (see [Console and Torasso, 1991], Section 6). Lenient explanations place a premium on causal explanations.

In summary, LIEs are generated by the following steps:

1. Find the lenient explanations of O .
2. Of these, choose the ones that have a minimal adjunct.
3. Of these, choose the ones that have minimal nonnormal causes. These are the lenient ideal explanations of O .

4 Computational methods

We develop some computational methods that can be applied to generate LIE's for an observation set. These methods are similar to the minimal conflict methods of diagnostic theories [de Kleer *et al.*, 1990; Reiter, 1987]. We use the XOR function circuit diagrammed in Figure 7 as an example. There are three gates, a , b , and c , with two inputs (i , j) and one output (o).

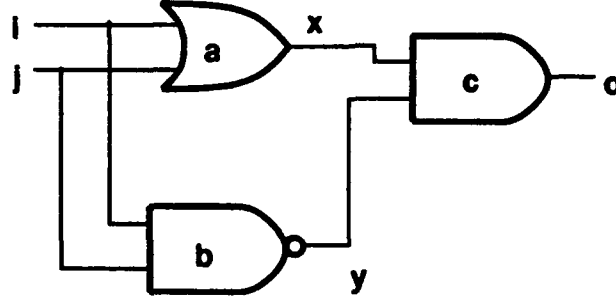


Figure 7: An XOR Circuit

The atoms i, j, o, x and y stand for circuit logic levels, so that i means input i is one, \bar{i} that it is zero. We have the following causal relation R :

$$\begin{array}{lll}
 i, ok(a) \supset x & x, y, ok(c) \supset o & i, j, ok(b) \supset \bar{y} \\
 j, ok(a) \supset x & \bar{x}, ok(c) \supset \bar{o} & \bar{i}, ok(b) \supset y \\
 \bar{i}, \bar{j}, ok(a) \supset \bar{x} & \bar{y}, ok(c) \supset \bar{o} & \bar{j}, ok(b) \supset y \\
 ab(a) \supset \bar{x} & ab(c) \supset \bar{o} & ab(b) \supset \bar{y}
 \end{array} \quad (13)$$

The only explicit fault mode is that when a gate fails, its output is stuck at zero. The normal conditions are $N = \{ok(a), ok(b), ok(c)\}$. The primitive causes are $C = \{i, \bar{i}, j, \bar{j}, ab(a), ab(b), ab(c)\}$.

4.1 Minimal conflicts and regular explanations

The first step is to consider compact ways to represent normal causal explanations. One idea is to just consider explanations that are subset minimal, which is a large reduction in the search space. Normal explanations are not minimal in this sense; nevertheless under certain circumstances we can represent all normal explanations as a combination of a minimal explanation and a maximally consistent set of normal conditions.

We will use only a single atom as the observation; an observation set O can be accommodated by using a new atom g , and adding the causal rule

$$o_1 \wedge \dots \wedge o_n \supset g,$$

and taking f as the single observation.

Recall that the adjunct of a explanation A for g is the set of normal conditions that are not exempted by A or whose complement is in A . The exemptions are a complicating factor, since they may introduce causes into A that have nothing to do with the derivation of g . Let us call an explanation A asserting no exemptions a *regular* explanation. We first develop methods for regular explanations.

The adjunct of a regular explanation A has every normal condition not contained in A . The adjunct of A is written as $\text{adj}(A)$. Let us define an *extension* of an explanation A as A together with a maximally consistent set of normal conditions. A *minimal explanation* A is one that is subset-minimal over all explanations.

It is useful to represent explanations by their essential elements, i.e., the ones that explain g .

Definition 7 Let A be an explanation for g . A core of A is a subset A' of A such that:

1. A' is a minimal explanation of g .
2. There is some extension E of A' such that $\text{adj}(E) = \text{adj}(A)$.

A core of A represents A in the sense that it can be extended to an explanation with the same adjunct. Given this, we can find all normal explanations of g (assuming they are regular) by first constructing a set of explanations Σ containing at least the core of every normal explanation, and then selecting the subset $\mu(\Sigma)$ whose elements have a minimal adjunct.

We first show that the set of minimal explanations covers the cores of all normal explanations. For regular, normal A , every minimal explanation contained in A is a core of A .

Fact 3 For every explanation A there exists a minimal explanation $A' \subseteq A$. If A is regular and normal, then A' is a core of A .

Proof. It is obvious that A embeds a minimal explanation. To find one, just keep discarding elements of A until the ones left are necessary for deriving g .

Let X be the set of normal conditions $(A - A') \cap N$. Assume that A is regular and normal; then A' is also regular. If $X \cup A'$ is an extension

of A' , then we are done, because $\text{adj}(X \cup A') = \text{adj}(A)$. So suppose the $X \cup A'$ is not normal, and there is another condition $n \in N$ that can be consistently added to it. But then A cannot be normal, because $\text{adj}(X \cup A' \cup \{n\}) \subset \text{adj}(A)$, contradicting the original hypothesis.

This result is encouraging. It suggests that we can find normal (regular) explanations for g by looking at its minimal explanations and comparing the adjuncts of their extensions.

Example 11 *We consider the XOR circuit with axioms (19), plus the additional fact that if c fails, so does one of a or b :*

$$ab(c) \supset (ab(a) \vee ab(b)) \in I.$$

Note that there will be no causal exemptions, because no ab predicate appears in the head of a clause of the causal relation. Thus all explanations will be regular.

Assuming \bar{j} and i as initial conditions, there are three minimal explanations of \bar{o} . We list these with the normal conditions of their extensions.

Minimal explanation	Extension
1. $ab(c)$	$ok(a)$ $ok(b)$
2. $ab(a), ok(c)$	$ok(c), ok(b)$
3. $ab(b), ok(c)$	$ok(c), ok(a)$

It is easy to check that the normal explanations are the extensions of 2 and 3.

Using Fact 3, we next show that all normal explanations can be found by comparing the adjuncts of minimal explanations.

Fact 4 *Suppose all the normal explanations of g are regular. Let Σ be the minimal explanations of g , and $\mu(\Sigma)$ the subset of Σ with minimal adjuncts. The elements of $\mu(\Sigma)$ are exactly the cores of all normal explanations of g .*

Proof. From Fact 3 we know that the core of every normal explanation must be in Σ . Since no explanation can have an adjunct properly contained in those of a normal explanation, every member of $\mu(\Sigma)$ must be the core of some normal explanation.

From this result, we need only compare the adjuncts of the minimal explanations for g in order to find the normal ones.

Example 12 *Considering again the XOR circuit with the added correlation of example 11, and assuming \bar{j} and i as initial conditions, we list the adjuncts for each of the three minimal explanations of \bar{o} .*

Minimal explanation	Extension	Adjunct
1. $ab(c)$	$ok(b)$	$ok(c), ok(a)$
	$ok(a)$	$ok(c), ok(b)$
2. $ab(a), ok(c)$	$ok(b), ok(c)$	$ok(a)$
3. $ab(b), ok(c)$	$ok(a), ok(c)$	$ok(b)$

Explanations 2 and 3 have minimal adjuncts, so they are the cores of the normal explanations of \bar{o} .

One way to compute the adjuncts of the minimal explanations of g is to use the method of minimal conflicts and candidates [de Kleer and Williams, 1987; Reiter, 1987]. If there are many normal conditions, minimal conflicts are usually a much more efficient means of finding the adjuncts of a minimal explanation than enumerating its extensions.

The definitions follow closely those of [de Kleer and Williams, 1987; Reiter, 1987], but are relativized to a given causal explanation. A *minimal conflict* for an explanation A is a minimal set of normal conditions that is inconsistent with $A \cup I$. A *candidate* for A is a minimal set that contains at least one element from each minimal conflict (candidates are called *hitting sets* in [Reiter, 1987]). The candidates of A can be generated from the minimal conflicts of A by picking one element from each minimal conflict. If A is regular, we can show that the candidates of A are just the adjuncts of the extensions of A .

Fact 5 *Let A be a regular explanation. X is a candidate of A if and only if it is the adjunct of some extension of A .*

Proof. Let X be a candidate of A , and let $\bar{X} = N - X$. We will show that $A \cup \bar{X}$ is an extension, that is, no more normal conditions can be consistently added to it. Suppose to the contrary that $x \in X$ is consistent with $A \cup \bar{X}$. There is a minimal conflict set S containing x but no other member of X (if not, $X - x$ would be a candidate, and X would not be). All the element of S except x are in $A \cup \bar{X}$; therefore x is inconsistent with it. Since $A \cup \bar{X}$ is an extension of A , and A and A' are regular, X is its adjunct by Fact 3.

In the other direction, let A' be an extension of A . We will show that $\text{adj}(A')$ is a candidate for A . Since A and A' are regular, $\text{adj}(A')$ is the set $Y = N - A'$ (Fact 3). Y must contain at least one element from each minimal conflict, otherwise it would be inconsistent. Suppose that Y is not a candidate. Then there is an element $y \in Y$ that is redundant, i.e., $Y - y$ contains a member of each minimal conflict. This y is consistent with A' , which means that A' cannot be an extension, contradicting the hypothesis.

Example 13 *We redo the last example using these techniques. Here are the minimal conflicts and their candidates for each of the three minimal explanations of \bar{o} .*

Minimal explanation	Conflicts	Candidate
1. $ab(c)$	$ok(a), ok(b)$	$ok(c), ok(a)$
	$ok(c)$	$ok(c), ok(b)$
2. $ab(a), ok(c)$	$ok(a)$	$ok(a)$
3. $ab(b), ok(c)$	$ok(b)$	$ok(b)$

The candidates are just the adjuncts of the minimal explanations. In general there can be candidates which are not adjuncts of minimal explanations, but these will always be subsumed by some other candidate that is. In this case the minimal conflict encoding is as complex as finding the extensions directly, but as the number of normal conditions gets larger, the minimal conflict encoding tends to be much more compact.

If the normal explanations of an observation are not regular, then the method of comparing adjuncts of the extensions of minimal explanations

will not identify them. In the non-regular case, we must instead look at an expanded class of explanations, rather than just the minimal ones of g .

Definition 8 (Active explanation)

Any abnormal condition (the complement of a normal condition) that has an explanation in a DCN is called an active condition. An active explanation for an observation g is a minimal explanation for $\{g\} \cup W$, where W is any set of active conditions.

That is, the active explanations for g are just the minimal explanations for g expanded by minimal explanations for some active conditions. We can now define an active core for an explanation.

Definition 9 *Let A be any explanation for g . An active core of A is a subset A' of A such that:*

1. A' is an active explanation of g .
2. There is some extension E of A' such that $\text{adj}(E) = \text{adj}(A)$.

We can show that the set of active explanations covers the active cores of all normal explanations. For normal A , there is an active explanation contained in A that is an active core of A .

Fact 6 *For every normal explanation A there exists a subset $A' \subseteq A$ such that A' is an active core of A .*

Proof. The proof is similar to that for Fact 3. Let X be the normal conditions contained in A , and Y the conditions exempted by A , so that $\text{adj}(A) = N - (X \cup Y)$. Let A' be a minimal subset of A such that the elements of Y are causally exempted by A' . Form the extension $E = A' \cup X$. Now $\text{adj}(E) \subseteq \text{adj}(A)$, since E contains all of X and causally exempts all of Y . If $\text{adj}(E) \subset \text{adj}(A)$, then A is not a normal explanation, contradicting the hypothesis.

Now we can use the same techniques as for regular explanations, namely, find all active explanations, and choose the ones with minimal adjuncts. These will be the active cores of the normal explanations. Since exempted conditions are explicitly explained by active explanations, the adjuncts can

be readily computed by using the minimal conflict method to find the normal conditions contained in the adjunct, and then subtracting the exempted conditions.

Example 14 Consider the basic XOR circuit, with the addition that whenever c is abnormal it causes b to be abnormal: $ab(c) \supset ab(b) \in R$. Assuming \bar{j} and i as initial conditions, we list the minimal conflicts and adjuncts for each of the active explanations of \bar{o} .

Active explanation	Conflicts	Adjunct
1. $ab(c)$	$ok(b)$ $ok(c)$	$ok(c)$
2. $ab(a), ok(c)$	$ok(a)$	$ok(a)$
3. $ab(b), ok(c)$	$ok(b)$	$ok(b)$

The active explanations in this case are the minimal explanations although in general they need not be. The adjuncts can be computed from the conflicts; the adjunct of 1 does not contain $ok(b)$, because its complement is caused by $ab(c)$, and so exempted. By comparing adjuncts we conclude that the extension of 1 (containing $ok(a)$) is normal, as well as the extensions of 2 and 3.

At this point we have enough results to form a proof method for finding the LIEs of an observation set O , assuming that the causal relation is finite, and all minimal conflicts are finite and computable.

1. Find the maximal subsets of O that have lenient explanations; call these the lenient subsets.
 \Rightarrow Find minimal causal explanations for all subsets of O , and check whether they are consistent with $O \cup I$. Choose the maximal subsets of O that have such explanations.
2. Find the active explanations for these subsets of O .
 \Rightarrow Adjoin to O all the possible subsets of the active conditions, and find all minimal explanations for each.

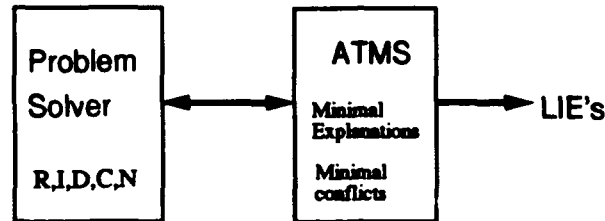


Figure 8: A Problem-Solving Architecture

3. Find the normal explanations among all explanations for the lenient subsets of O .
 \Rightarrow Find the conflicts of the active explanations of the previous step, and generate the adjuncts. The normal explanations are the extensions of the active explanations with minimal active candidates.
4. The ideal explanations are the normal explanations with minimal non-normal causes.

4.2 ATMS implementation

The proof method just given can be implemented using a modification of the computational techniques available in the ATMS. Because the full language of the correlational theory I of DCNs is first-order, and the causal relation may be infinite, it is not possible in general to have a complete proof theory (as is also the case for normal default logic). Instead, the computation of LIEs can be phrased in terms of a dialogue between a problem solver and the ATMS [de Kleer, 1986], as in Figure 4.2.

The problem solver is an inference engine containing the DCN theory. It computes two kinds of structures and sends them to the ATMS. The definite clause causation relation is sent directly as an ATMS definite clause. These clauses are stored by the ATMS and used to compute the minimal explanations for all literals c present in the ATMS.

The ATMS keeps track of inconsistent sets of causes and normal conditions through the use of its NOGOOD mechanism. Whenever the problem solver finds a set of literals whose conjunction is inconsistent with I , it can send them to the ATMS as a NOGOOD. The *label* of a node g in the ATMS is the set of all minimal $A \subseteq C \cup N$ such that $A \models_R g$ and A is consistent

with the NOGOODS. If the ATMS has the complete relation R , and the NOGOODS completely cover the inconsistent causes, then the label of g is the set of minimal explanations for g .

We need two additional structures: an observation set node, and an active condition set node. LIEs are found by computing the candidates of the labels of the conjunction of these two nodes.

We have successfully tried all of the problems in this paper. Because the ATMS algorithms are exponential, there can be difficulties in scaling up to larger problems. The techniques developed for using heuristics in the ATMS might help here, and the causal relation could be used to focus the work of the ATMS. However, currently we have no experience with large problems.

5 Relation to other approaches

We have already critiqued the consistency-based approach to MBD in Section 2. More recently, several abductive approaches have been developed, among them [Console and Torasso, 1991; Poole, 1989; Poole, 1993; Dressler and Struss, 1992; Besnard and Cordier, 1993]. These methods are similar to DCNs in their use of abduction to explain rather than excuse observations. For MBD, abductive explanations are typically defined as follows.

Definition 10 *Let $(SD, CMPS)$ be a system, and let Δ be a subset of $CMPS$. Define $\Delta_{ab} = \{ab(c) \mid c \in \Delta\} \cup \{\neg ab(c) \mid c \in CMPS - \Delta\}$. Then Δ is an abductive diagnosis of the observations O if it is a minimal set such that*

1. $SD \cup \Delta_{ab} \vdash O$ and
2. $SD \cup \Delta_{ab} \cup O$ is consistent.

This is the same definition as for the consistency-based approach (Definition 1), with the addition of the first clause stating the the observations must follow from the system description. It is also similar to the definition of explanation in DCNs (3), but there is no distinction between the causation relation and correlation.

The general relation between abductive and consistency-based approaches to MBD is pointed out in [Poole, 1988a], [Console *et al.*, 1988], and [Konolige, 1992]. The type of information needed is different: in the abductive method,

one uses forward-working axioms to derive observations from component behavior, e.g., the implications of Equation (1) and the fault model (4). Note that the abductive method can use partial fault models. These are typically expressed as behavioral modes of the components, i.e., axioms of the form

$$\begin{aligned} \text{fault}_i(c) &\supset P_i(c) \\ \text{fault}_i(c) &\supset \text{ab}(c) \end{aligned} \quad (14)$$

Each $\text{fault}_i(c)$ is a particular way in which the component can fail, and $P_i(c)$ describes that failure. For the consistency-based method to return similar diagnoses, the system description must be augmented by closure axioms stating that the only faults are those explicitly given, that is, axioms of the form

$$\text{ab}(x) \supset \text{fault}_1 \vee \dots \vee \text{fault}_n. \quad (15)$$

The abductive method also solves the problem of relevance, in the sense that the diagnosis Δ involves components that, if abnormal, imply the observations. And it can accommodate neutral causes, by making them hypotheses, but not classifying them as either normal or abnormal. But, there are still several ways in which the abductive approaches differ from DCNs.

- There is no explicit causal relation. The explanations given are not causal implications, and the problems noted in Section 2.3 apply.
- The definition of abductive diagnosis is complicated by the presence of both normality and fault assumptions. In the theory of DCNs, it was the interaction of these two that produced the complications of normal and ideal explanations. The abductive approaches to MBD have problems in formulating a parsimony criteria for explanations.

We examine two representative examples of the abductive approach in more detail: Poole's THEORIST system, and Console, Dupré and Torasso's merging of consistency-based and abductive approaches.

5.1 Poole's THEORIST

Poole [Poole, 1989; Poole, 1993] develops an abductive approach to diagnosis using his THEORIST system. Given a system (SD, CMPS), a normality assumption is a predicate $\neg \text{ab}(c)$ for some component c , and a fault assumption is a predicate $\text{fault}_i(c)$ such that $\text{fault}_i(c) \supset \text{ab}(c)$.

Definition 11 Let D be a set of normality and fault assumptions. A diagnosis for observations O is a minimal set D such that

1. $SD \cup D \vdash O$ and
2. $SD \cup D \cup O$ is consistent.

This definition is similar to the general abductive framework of Definition 10, with two differences. It makes the explicit assumption of behavioral modes for faults; this is one of the first logical frameworks to incorporate fault models along with the normal behavior of the system. Second, D is a minimal set of normality and fault assumptions together. From the DCN point of view, Poole's framework does not enforce maximal normality (even without causal precedence), and so will generate spurious diagnoses. On the other hand, Poole's system does have a strong parsimony criterion for both normal conditions and faults, giving only the assumptions that are strictly necessary for predicting the observations. Finally, other aspects of DCNs, such as the integration of excuses and explanations into lenient explanations, are not present in Poole's framework.

5.2 Console and Torasso

Console and Torasso [Console and Torasso, 1991] generalize both the abductive and consistency-based approaches. Like Poole, they use a system description that describes normal functioning of the system, as well as behavioral fault modes. They define a *diagnostic problem* as a system description, a set of components, observations to be explained, and a context for the explanation. The context is a set of conditions that are observed or hypothesized but do not need explanation. They can serve as initial conditions to the diagnosis problem, or as hypotheses for neutral causes. For our purposes, we consider the context to be null.

Diagnoses are constructed from complete assignments of behaviors to the components.

Definition 12 Let W be a set composed of $ab(c)$ or $fault_i(c)$ for each component c , and O^+ a subset of O . A diagnosis is a set W such that

1. $SD \cup W \vdash O^+$ and
2. $SD \cup W \cup O$ is consistent.

The subset O^+ is the part of the observations that are explained abductively. By varying O^+ , the definition can be made to range from consistency-based ($O^+ = \emptyset$) to fully abductive ($O^+ = O$). This behavior is similar to that of lenient explanations in DCNs, where a maximal subset of the observations are causally explained, and the rest are subject to excuses.

The definition of diagnosis here differs from Poole's in that a complete set of behaviors for components are assumed. In Console and Torasso's system, a diagnosis that is minimal in fault assumptions will have a maximal set of normal assumptions, unlike in Poole's system, in which both normal and fault assumptions are minimized. This corresponds to normal explanations in DCNs, although of course there is no causal exemption.

On the other hand, Poole's system is better in preferring explanations with a minimal set of faults. Consider the following example:¹

$$\begin{aligned} ab(1) &\supset g \\ c, ab(2) &\supset g \\ ab(1) &\supset ab(2) \end{aligned} \tag{16}$$

To explain g given the initial condition c , Poole's system will give the two minimal sets $\{ab(1)\}$ and $\{ab(2)\}$. This seems intuitively correct, since either of the explanations would be equally likely, given equal priors for $ab(1)$ and $ab(2)$. Console and Torasso will have two explanations, $\{ok(1), ab(2)\}$ and $\{ab(1), ab(2)\}$. The first of these is minimal in faults, and so would be preferred.

Console and Torasso also mention preferences based on implication, i.e., explanation E_1 is preferred to E_2 if $E_2 \models E_1$ but not *vice versa*. This is a way of producing partial explanations from the complete sets W . However, this method treats normality and fault assumptions equally, whereas the selection of explanations in DCNs is a two-step process, first maximizing normality assumptions and then minimizing abnormalities.

¹In fact, Console and Torasso explicitly forbid the presence of abnormality predicates in the head of a clause, ruling out the following example. They do not state why this is the case, and in the concluding section they relax this assumption.

5.3 Other approaches to explanation

Although we have concentrated on the application of DCNs to diagnosis, they provide a general framework for representing causation and explanation. Causation can be used as a unifying concept to understand various perspectives on diagnosis: excusing vs. explaining observations, correlation vs. causation, and the integration of normal conditions with explanatory causes. Although many of these issues have been dealt with separately in the literature, there have been few attempts to draw them together into a single framework, and the issues are often obscured by the formal or computational paradigm. There are many formal nonmonotonic systems that provide similar capabilities, although they are not phrased in terms of causation, e.g., Poole's THEORIST [Poole, 1988b]. DCNs are distinguished by providing a coherent account of causation, correlation, and default conditions. Perhaps the closest system is Geffner's theory of causal and conditional reasoning [Geffner, 1989], which also takes causation as a primitive concept, and ties together explanation, defaults, and causation. He provides a complex but plausible formal account of these concepts, using a modal expression $C\alpha$ to represent " α is caused." Although the formalisms differ, there are many points of similarity between this work and his. Perhaps the major difference is that the roots of DCNs are default logic and abductive inference, and thus there are natural computational methods using the ATMS.

A good test of the DCN framework is the application to reasoning about events. We have started this task, and it appears that the problems of causation, explanation, and prediction in an event calculus can be treated within the DCN framework. The approach is similar to that of Shanahan [Shanahan, 1989], but the formal machinery is more general, and includes causation.

6 Some remarks about causation

Perhaps the weakest point of the DCN approach is that the theory of causation is not well developed. Since causation is treated as a proof-theoretic concept, there are some obvious problems (or, one might say opportunities) that arise. We discuss some of these here; a more detailed treatment can be found in [Konolige, 1991].

and cumulative:

If $A \vdash_R c$ and $B, c \vdash_R d$, then $A, B \vdash_R d$.

As we have stated, the important part of the causal relation is that it captures the functional dependence of the domain variables; this is the main difference between a causal relation and a merely correlational one. The asymmetry of causation is represented by the asymmetry of inference in a definite clause theory.

These remarks leave open the question of whether, in a particular instance, it is possible to have a causation relation that is symmetric for two propositions, or more generally to have one that is cyclic, containing a loop that leads from a proposition back to the same proposition. Other commitments may answer this question: for instance, assuming that causes always precede their effects in time forces the causal relation to be acyclic. The definite clause theory itself does not enforce any acyclic condition.

There are some further complications in defining a causal relation that we will mention here, without offering any definitive solutions. The first is that of inferred causation. We mentioned this briefly in proposing the definitional theory in Section 3.2. We use only a simple form of definitions to represent complements; any full-fledged theory of causation should at least take into account abstraction relations among propositions, e.g., "A 40-watt bulb is a type of bulb."

Another problem arises when our knowledge of the causation relation is partial. We have already remarked that we may only know a subset of the actual causation relation. Other kinds of uncertainty also exist. For example, suppose we know that dialing the number "911" connects one with either the police or the fire department, but we don't know which. The action of dialing 911 is completely determinate, it's just that we don't know the exact outcome. To express epistemic uncertainty of this kind, it is necessary to describe the causation relation in an appropriate language. If we let c stand for the action of dialing 911, d for calling the police, and e for calling the fire department, then our knowledge is expressed by the statement:

Either $c \vdash_R d$ or $c \vdash_R e$.

DCNs are not expressive enough to state this; a language that talks about causation, such as Geffner's [Geffner, 1989], would be necessary.

First, there is a deliberate sloppiness about stating propositions in the causal relation. Most of the ones used in this paper are statements about particular properties, e.g., the switch is closed or the light is on. But causation also involves events: "closing the switch caused the light to go on." We are trying to be as noncommittal as possible about the ontology of events and propositions, whether states of the world can be allowed as causes, how to specify the time of events, and so on. Any consistent defensible set of choices will do.

The second point is that a definite clause of R must specify *all* and *only* the propositions governing an effect. Closing the switch only turns on the bulbs if they are ok and the wires are intact. Of course, in any real-world situation there will be an inordinate number of such conditions, so any default causal theory will be relative to a set of background assumptions that do not enter into the theory. The choice of these assumptions is conventional.

It is important that only the relevant propositions participate in the causal relation. If we add an irrelevant proposition to the antecedent of a clause, the relation would still be useful in the sense that conjunction of the antecedents produces the desired effect, but it would be misleading in implying that all the antecedents were necessary. In producing explanations, minimal causal antecedents are required in the causal relation to ensure that explanations do not contain irrelevant propositions.

The role of primitive causes is to define the propositions over which, in some sense, we can exercise direct control. The point at which we choose to define primitive causes is partly a matter of convention. Often bodily movements are taken to be the ultimate primitive causes, but this viewpoint is unnecessarily restrictive. Any well-defined event or condition that we can reliably bring about will suffice for a primitive cause, as long as the purpose of producing explanations is to give a set of conditions that account for the observed facts, and over which we have control.

One way to understand the causation relation R is as a provability relation. The provability relation is composed from individual inference steps combined into a tree; in the same way, the causation relation is specified by combining definite clause inference steps into a proof. Like classical provability, causation is monotonic:

$$\text{If } A \vdash_R c \text{ and } B \supset A, \text{ then } B \vdash_R c$$

7 Conclusion

We have developed a theory of causation in the presence of defaults about normally occurring conditions. The theory is based on a structure called Default Causal Nets, which integrate causal, correlational, and definitional information. These nets can be used to generate predictions and explain observations.

We have argued that preferences among explanations can be based on noting how causation and defaults interact, as in Examples 1 and 2. Such preferences seem to follow commonsense reasoning based on causal knowledge. In model-based diagnosis, any assumptions about causation and defaults are implicit in the representation of components as being normal or abnormal, and the search for diagnoses is based on abnormal components. Such a view, we argue, is representationally restrictive, and does not give a deep enough analysis about how defaults interact. For example, although we can state relations among abnormalities in the domain, these relations do not necessarily lead to intuitively correct preferences among diagnoses in the consistency-based approach, because material implications within the framework are not treated as causal relations.

Acknowledgments

The research reported in this paper was supported by the Office of Naval Research under Contract No. N00014-89-C-0095.

I would like to thank Luca Console, Oskar Dressler, Gerhard Friedrich, Hector Geffner, Moises Goldszmidt, Ilkka Niemelä, and Peter Struss for valuable discussions.

References

- [Besnard and Cordier, 1993] Philippe Besnard and Marie-Odile Cordier. Explanatory diagnoses and their computation by circumscription. *Annals of Mathematics and Artificial Intelligence*, XX, 1993.
- [Cohen et al., 1990] P. R. Cohen, J. Morgan, and M. E. Pollack, editors. *Intentions in Communication*, Cambridge, MA, 1990. MIT Press.

- [Console and Torasso, 1991] L. Console and P. Torasso. A spectrum of logical definitions of model-based diagnosis. *Computational Intelligence*, 7(3):133-141, 1991.
- [Console *et al.*, 1988] L. Console, D. Theseider Dupre, and P. Torasso. Abductive reasoning through direct deduction from completed domain models. In Z. W. Ras, editor, *Methodologies for Intelligent Systems 4*, pages 175-182. North-Holland, 1988.
- [Davis and Hamscher, 1988] R. Davis and W. Hamscher. Model-based reasoning: Troubleshooting. In H. E. Shrobe, editor, *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence*, pages 297-346. Morgan Kaufmann, San Mateo, CA, 1988.
- [de Kleer and Williams, 1987] Johan de Kleer and Brian C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97-130, 1987.
- [de Kleer and Williams, 1989] Johan de Kleer and Brian C. Williams. Diagnosis with behavioral modes. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989.
- [de Kleer *et al.*, 1990] Johan de Kleer, Alan Mackworth, and Ray Reiter. Characterizing diagnoses. In *Proceedings of the Conference of the American Association of Artificial Intelligence*, Boston, MA, 1990.
- [de Kleer, 1986] Johan de Kleer. An assumption-based truth maintenance system. *Artificial Intelligence*, 28:127-162, 1986.
- [Dressler and Struss, 1992] O. Dressler and P. Struss. Back to defaults: Characterizing and computing diagnoses as coherent assumption sets. In *European Conference on Artificial Intelligence*, Vienna, 1992.
- [Friedrich *et al.*, 1990] Gerhard Friedrich, Georg Gotlob, and Wolfgang Nejdl. Physical impossibility instead of fault models. In *Proceedings of the Conference of the American Association of Artificial Intelligence*. MIT Press, 1990.
- [Geffner, 1989] Hector Geffner. *Default Reasoning: Causal and Conditional Theories*. PhD thesis, Department of Computer Science, University of California at Los Angeles, 1989.

- [Junker, 1993] Ulrich Junker. Preferring diagnoses using a partial order on assumptions. *Annals of Mathematics and Artificial Intelligence*, XX, 1993.
- [Konolige, 1991] Kurt Konolige. What's happening: elements of common-sense causation. In *Proceedings of the International Conference on Cognitive Science*, San Sebastian, Spain, May 1991.
- [Konolige, 1992] Kurt Konolige. Abduction vs. closure in causal theories. *Artificial Intelligence*, 53(2-3), 1992.
- [Lewis, 1973] D. Lewis. *Counterfactuals*. Blackwell, 1973.
- [Nayak, 1992a] P. Pandurang Nayak. Causal approximation. In *Proceedings of the Conference of the American Association of Artificial Intelligence*, pages 703-709, Menlo Park, CA, 1992. AAAI Press/MIT Press.
- [Nayak, 1992b] P. Pandurang Nayak. Order of magnitude reasoning using logarithms. In *Proceedings of the International Conference on Knowledge Representation and Reasoning*, pages 201-210, San Mateo, CA, 1992. Morgan Kaufmann.
- [Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
- [Poole, 1988a] D. Poole. Representing knowledge for logic-based diagnosis. In *Proceedings of the International Conference on Fifth Generation Computing Systems*, pages 1282-1290, Tokyo, 1988.
- [Poole, 1988b] David Poole. A methodology for using a default and abductive reasoning system. Technical report, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1988.
- [Poole, 1989] David Poole. Normality and faults in logic-based diagnosis. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989.
- [Poole, 1993] David Poole. Representing diagnosis knowledge. *Annals of Mathematics and Artificial Intelligence*, XX, 1993.

- [Reggia *et al.*, 1985] J. A. Reggia, D. S. Nau, and Y. Wang. A formal model of diagnostic inference I. Problem formulation and decomposition. *Inf. Sci.*, 37, 1985.
- [Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32, 1987.
- [Shanahan, 1989] Murray Shanahan. Prediction is deduction but explanation is abduction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989.
- [Shoham, 1987] Yoav Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, Massachusetss, 1987.
- [Struss and Dressler, 1989] P. Struss and O. Dressler. Physical negation – integrating fault models into the general diagnostic engine. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1318–1323, Detroit, MI, 1989.
- [Suppes, 1970] P. Suppes. *A probabilistic theory of causation*. North Holland, Amsterdam, 1970.

A Representationalist Theory of Intention

Kurt Konolige*

Artificial Intelligence Center
SRI International
Menlo Park, CA 94025

Martha E. Pollack†

Dept. of Computer Science
Univ. of Pittsburgh
Pittsburgh, PA 15260

Abstract

Several formalizations of cognitive state that include intentions and beliefs based on normal modal logics (NMLs) have appeared in the recent literature. We argue that NMLs are not an appropriate representation for intention, and provide an alternative model, one that is representationalist, in the sense that its semantic objects provide a more direct representation of cognitive state of the intending agent. We argue that this approach results in a much simpler model of intention than does the use of an NML, and that, moreover, it allows us to capture interesting properties of intention that have not been addressed in previous work.

1 Introduction

Formalizations of cognitive state that include intentions and beliefs have appeared in the recent literature [Cohen and Levesque, 1990a; Rao and Georgeff, 1991; Shoham, 1990; Konolige and Pollack, 1989]. With the exception of the current authors, these have all employed *normal modal logics* (NMLs), that is, logics in which the semantics of the modal operators is defined by accessibility relations over possible worlds. This is not surprising, since NMLs have proven to be a powerful tool for modeling the cognitive attitudes of belief and knowledge. However, we argue that intention and belief are very different beasts, and that NMLs are ill-suited to a formal theory of intention.

We therefore present an alternative model of intention, one that is representationalist, in the sense that its semantic objects provide a more direct representation of cognitive state of the intending agent. We argue that this approach results in a much simpler model of intention than does the use of an NML, and that, moreover, it allows us to capture interesting properties of intention that have not been addressed in previous work. Further,

the relation between belief and intention is mediated by the fundamental structure of the semantics, and is independent of any particular choice for temporal operators or theory of action. This gives us a very direct, simple, and semantically motivated theory, and one that can be conjoined with whatever temporal theory is appropriate for a given task.

In the next section (Section 2), we make the case for a representationalist theory of intention. Section 3 constitutes the technical heart of our paper: there we develop our formal model of intention. Finally, in Section 4, we draw some conclusions and point the way toward further development of our logic of intention.

2 The case for representationalism

As we noted above, NMLs have been widely and successfully used in the formalization of belief. It is largely as a result of this success that researchers have adopted them in building models of intention. However, we argue in this section that these logics are inappropriate to models of intention:

- The semantic rule for normal modal operators is the wrong interpretation for intention. This rule leads to the confusion of an intention to do ϕ with an intention to do any logical consequence of ϕ , called the *side-effect problem* [Bratman, 1987]. A simple and intuitively justifiable change in the semantic rule makes intention side-effect free (and nonnormal).
- Normal modal logics do not provide a means of relating intentions to one another. Relations among intentions are necessary to describe the means-end connection between intentions.

NMLs are closed under logical consequence: given a normal modal operator L , if $L\phi$ is true, and $\phi \models \psi$, it follows that $L\psi$ is true. When L represents belief, consequential closure can be taken to be an idealization: although it is obviously unrealistic in general to assume that an agent believes all the consequences of his beliefs, it is reasonable to assume this property of an ideal agent, and this idealization is acceptable in many instances.

However, consequential closure *cannot* be assumed for intention, even as an idealization. It is clear that an agent who intends to perform an action usually does not intend all the consequences of that action, or even all the

*Supported by the Office of Naval Research under Contract No. N00014-89-C-0095.

†Supported in part by a National Science Foundation Young Investigator's Award (IRI-9258392), by the Air Force Office of Scientific Research (Contract F49620-92-J-0422), and by DARPA (Contract F30602-93-C-0038).

consequences he anticipates. Some of the consequences are goals of the agent, while others are "side effects" that the agent is not committed to.¹

Because NMLs are subject to consequential closure, and intention is not, several strategies are used to make the logics side-effect free. They all involve relativizing the side-effects of intentions to believed consequences. The thesis of *realism* is that all of an agent's intended worlds are also belief worlds [Cohen and Levesque, 1990a], that is, a rational agent will not intend worlds that he believes are not possible. Given the realism thesis, whenever the agent intends a and believes $a \supset b$, he will also intend b . Cohen and Levesque [Cohen and Levesque, 1990b] adopt the realism thesis, and rely on claims about way an agent may change his beliefs about the connection between an intended proposition and its consequences to make their theory side-effect free. In their case, an agent who always believes that $a \supset b$ is always true will incur the side-effect problem when intending a . Also, any analytic implication (i.e., when $a \supset b$ must be true in all possible futures) will cause problems. Two special cases are abstractions (e.g., making a dinner is an abstraction of making a spaghetti dinner) and conjunctions (intending $a \wedge b$ implies intending a and intending b separately).

Rao and Georgeff [Rao and Georgeff, 1991] point out that by relaxing realism, intentions can be made side-effect free. *Weak realism* is the thesis that at least one intended world is a belief world. There can thus be intention worlds that are not belief worlds. Now, even though the agent believes $a \supset b$, b is not an intention, because there is an intended world in which a is true but not b . Weak realism seems inherently less desirable than realism (how is it possible for an agent to intend worlds he does not believe possible?), and it is still not fully side-effect free, since it is closed under conjunctions and abstractions.

These problems do not mean we have to abandon possible worlds. In fact, with the right semantics, possible worlds are an intuitively satisfying way of representing future possibility and intention for an agent. We note that intentions divide the possible futures into those that the agent wants or prefers, and those he does not. Consider the diagram of Figure 1. The rectangle represents the set of possible worlds W . The *scenario* for a proposition a is the set of worlds in W that make a true: the shaded area in the diagram. An agent that has a as an intention will be content if the actual world is any one of those in the shaded area, and will be unhappy if it is any unshaded one. The division between wanted and unwanted worlds is the important concept behind scenarios. For example, consider another proposition b that is implied by a (for concreteness, take a = "I get my tooth filled," and b = "I feel pain.") If we just look

¹For example, an agent may intend to go to the dentist to get his tooth filled, believing that he will feel pain as a consequence, without being committed to feeling the pain. If he discovers that the dental work is painless, he will not seek to experience the pain nonetheless. See Bratman [Bratman, 1987] and Cohen and Levesque [Cohen and Levesque, 1990b] for further discussion.

Possible worlds W

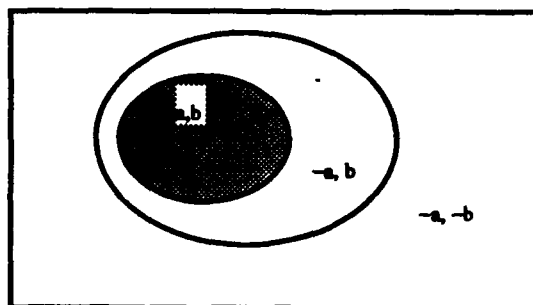


Figure 1: A Venn diagram of two scenarios.

at interpretations within the shaded area, a and b both hold, and so cannot be distinguished. But the complement of these two propositions is different. A world in the area $\neg a, b$, in which the agent feels pain but does not have his tooth pulled, is an acceptable world for the intention b , but not for a . So the interpretation rule for intention must take into account the complement of the intended worlds. As we will see in Section 3, this makes intention a nonnormal modal operator. It also makes it side-effect, abstraction, and conjunction free, whether we choose realism or weak realism.

The representationalist part of the model comes in representing the mental state of the agent using scenarios. *Cognitive structures*, containing elements representing intentions and the relationship among intentions, are used for this purpose.

3 Cognitive structures

Our model of intention will have two components: possible worlds that represent possible future courses of events, and *cognitive structures*, a representation of the mental state components of an agent. We introduce complications of the model in successive sections. To begin, we define the simplest model, a static representation of primary or "top-level" intentions. Primary intentions do not depend on any other intentions that the agent currently has.²

3.1 Possible Futures

The concept of intention is intimately connected with choosing among course of future action. In the model, courses of action are represented by possible worlds. Each possible world is a complete history, specifying states of the world at all instants of time. We assume there is a distinguished moment *now* in all worlds that

²This is a bit of an overstatement, since an agent's intentions change over time, and an intention that begins life as primary may later also be used in support of some other intention. In such cases we say that the intention has been *overloaded*. Overloading is a cognitively efficient strategy for an agent to employ [Pollack, 1991]. For the moment, however, we will not worry about primary intentions that later are overloaded.

is the evaluation point for statements.³

The set of possible worlds is W . For each world $w \in W$, there is an evaluation function that determines the value of sentences in a language \mathcal{L} , which refer to states of the world or actions that take place between states of this world. For any sentence ϕ of \mathcal{L} , $w(\phi)$ is the truth-value of ϕ .

To talk about contingent and necessary facts \mathcal{L} is extended to \mathcal{L}_\square , which includes the modal operators \square and \diamond . The possibility operator \diamond expresses the existence of a world with a given property. $\diamond\phi$ says that there is a world (among W) for which ϕ is true. Its semantics is:

Definition 3.1

$$w, W \models \diamond\phi \text{ iff } \exists w' \in W. w', W \models \phi.$$

\diamond is used to specify the background of physically possible worlds under which reasoning about intention takes place, and will be important in describing the structure of a given domain. The necessity operator $\square\phi$ is defined as $\neg\diamond\neg\phi$.

3.2 Belief and primary intentions

We begin by defining concept of scenario.

Definition 3.2 Let W be a set of possible worlds, and ϕ any sentence of \mathcal{L} . A scenario for ϕ is the set

$$M_\phi = \{w \in W \mid w, W \models \phi\}.$$

A scenario for ϕ identifies ϕ with the subset of W that make ϕ true.

A cognitive structure consists of the background set of worlds, and the beliefs and intentions of an agent.⁴

Definition 3.3 A cognitive structure is a tuple $\langle W, \Sigma, I \rangle$ consisting of a set of possible worlds W , a subset of W (Σ , the beliefs of the agent) and a set of scenarios over W (I , the intentions of the agent).

We extend \mathcal{L}_\square to \mathcal{L}_I by adding the modal operators B for belief and I for intentions. The beliefs of an agent are taken to be the sentences true in all worlds of Σ .⁵ For simplicity, we often write Σ as a set of sentences of \mathcal{L}_\square , so that M_Σ is the corresponding possible worlds set.

Definition 3.4

$$\langle W, \Sigma, I \rangle \models B(\phi) \text{ iff } \forall w' \in M_\Sigma. w', W \models \phi, \text{ i.e., } M_\Sigma \subseteq M_\phi.$$

³This definition of possible worlds is the one usually used in the philosophical literature, but differs from that of Moore in [Moore, 1980], where possible worlds are identified with states at a particular time.

⁴In this paper, we deal only with the single agent case, and thus we neither explicitly indicate the name of the (unambiguous) agent associated with any cognitive structure, nor include an agent argument in our intention or belief predicates.

⁵This enforces the condition of *logical omniscience* [Levesque, 1984] on the agent's beliefs, which is not a realistic assumption. We could chose a different form for beliefs, say a set of sentences of \mathcal{L}_I that is not closed with respect to consequence; but it would obscure the subsequent analysis.

The beliefs of an agent are always possible, that is, they are a subset of the possible worlds. This also means that an agent cannot be wrong about necessary truths. A more complicated theory would distinguish an agent's beliefs about what is possible from what is actually possible. The key concept is that intentions are represented with respect to a background of beliefs about possible courses of events (represented by \diamond), as well as beliefs about contingent facts (represented by B). Stated in \mathcal{L}_I , the following are theorems:

$$\begin{aligned} B(\phi) &\supset \diamond\phi \\ B(\square\phi) &\equiv \square\phi \end{aligned} \quad (1)$$

Of course, beliefs about contingent facts can still be false, since the real world does not have to be among the believed ones. The B operator represents all futures the agent believes might occur, including those in which he performs various actions or those in which he does nothing. The beliefs form a background of all the possibilities among which the agent can choose by acting in particular ways.

The third component of a cognitive structure for an agent, an intention structure, is a set of scenarios M_ϕ . Intuitively, an agent's intention structure will include one scenario for each of his primary intentions. We write I as a set of sentences of \mathcal{L}_\square , where each sentence ϕ stands for its scenario M_ϕ .

Definition 3.5

$$\langle W, \Sigma, I \rangle \models I(\phi) \text{ iff } \exists \psi \in I \text{ such that } M_\psi \text{ is a scenario for } \phi, \text{ i.e., } M_\psi = M_\phi.$$

This definition bears an interesting relation to the semantics of normal modal operators. Each primary intention (i.e., each element of I) acts like a separate modal operator. A normal modal operator I_ψ for the element M_ψ would be defined using:

$$\langle W, \Sigma, I \rangle \models I_\psi(\phi) \text{ iff } M_\psi \subseteq M_\phi,$$

just as for belief. The semantic rule for I is similar, but uses equality between the scenarios instead of subset, so that the worlds *not* in M_ψ must satisfy $\neg\phi$. By identifying intentions with scenarios, we explicitly encode in the semantics the distinction between preferred and rejected possible worlds. If we were to use the weaker form of the semantic rule for $I(\phi)$ (i.e., $M_\psi \subseteq M_\phi$), then there could be some world w which satisfies ϕ but is not a world satisfying the agent's intention. This is contrary to our reading of intention as a preference criterion dividing possible worlds.⁶

From this formal definition, it is easy to show that $I(\phi)$ will hold just in case ϕ is equivalent to some proposition $\psi \in I$, given the background structure W .

Proposition 3.1 For any structure $\langle W, \Sigma, I \rangle$,

$$\langle W, \Sigma, I \rangle \models I(\phi) \text{ iff } \exists \psi \in I. W \models \square(\phi \equiv \psi).$$

⁶Our semantics is also equivalent to the minimal model semantics of Chellas [Chellas, 1980]. In the minimal model semantics, the accessibility relation is from a world to a set of sets of worlds, i.e., a set of propositions. As Chellas shows, such logics are nonnormal, and the simplest system, E , contains only the inference rule $\phi \equiv \psi / I\phi \equiv I\psi$.

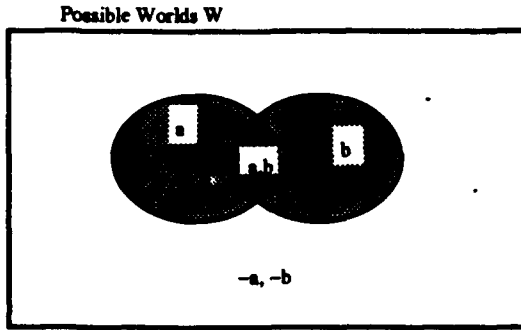


Figure 2: A Venn diagram of conjunctive scenarios.

The I operator is true precisely of the individual top-level intentions the agent has. It is not subject to closure under logical consequence or under the agent's beliefs. To see this, consider the cognitive structure $\langle W, \Sigma, \{a\} \rangle$, i.e., the agent has the single intention to perform a . Assume that a logically implies b , but not the converse, i.e.,

$$W \models \Box(a \supset b) \wedge \Diamond(b \wedge \neg a).$$

Then $M_a \neq M_b$, because there is a world in which b is true but a is not. From the semantics of I , we have

$$\langle W, \Sigma, \{a\} \rangle \models I(a) \wedge \neg I(b)$$

This shows that I is not closed with respect to valid consequence. To distinguish the intention of a from its necessary consequence b , there must be at least one possible world in which b is true but a is not. As a particular instance of this, our theory does not equate an intention to perform a conjunction with a conjunction of intentions. Assume that the set of possible worlds distinguishes a and b , i.e., $W \models \Diamond(a \wedge \neg b) \wedge \Diamond(\neg a \wedge b)$. Now consider two agents: the first has the single primary intention $a \wedge b$, and the second has exactly the two primary intentions a and b . Then:

$$\begin{aligned} \langle W, \Sigma, \{a \wedge b\} \rangle &\models I(a \wedge b) \wedge \neg I(a) \wedge \neg I(b) \\ \langle W, \Sigma, \{a, b\} \rangle &\models I(a) \wedge I(b) \wedge \neg I(a \wedge b) \end{aligned} \quad (2)$$

The reason for this is clear from the diagram of Figure 2. The scenario $M_{a \wedge b}$ excludes all interpretations outside of the overlap area in the figure; hence it is not equivalent to M_a , for which a perfectly acceptable world could contain a and $\neg b$; nor is it equivalent to M_b .

On the other hand, taking the two scenarios M_a and M_b singly, acceptable worlds are in the respective regions a and b . Thus the most acceptable worlds are in the overlap region. However, if one of the goals becomes impossible, say a , then any world in b is acceptable, unlike the case with the conjunctive scenario $M_{a \wedge b}$.

A similar story can be told for side effects and abstraction. The ability to distinguish between an intention and its side effects, abstractions, and conjunctions is basic to the semantics given in Definition 3.5, and does not require any further axioms or stipulations, nor any commitment to a particular temporal logic.

An alternative to the reading of "intention" as separate primary intentions is the reading as conjoined intention, i.e., " ϕ is intended if it is the intersection of worlds

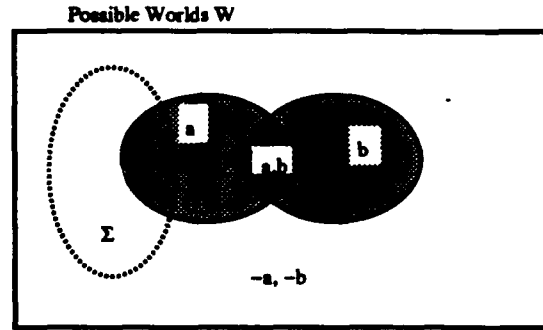


Figure 3: A Venn diagram of belief and intention.

of some set of primary intentions." We use the operator $I^*(\phi)$ for this reading.

Definition 3.6

$$\langle W, \Sigma, I \rangle \models I^*(\phi) \text{ iff } \exists J \subseteq I \text{ such that } M_J \text{ is a scenario for } \phi, \text{ i.e., } M_J = M_\phi.$$

I^* can be characterized by the following axioms.

Proposition 3.2 *The following are theorems of \mathcal{L}_I .*

$$\begin{aligned} I(a) &\supset I^*(a) \\ I^*(a) \wedge I^*(b) &\supset I^*(a \wedge b) \end{aligned}$$

So I^* sanctions the conjoining of separate intentions, but not the splitting of an intention that is a conjunction.⁷

3.3 Rationality constraints: intention and belief

So far we have not related the agent's intentions to his beliefs. Consider the diagram of Figure 3, for which the cognitive structure is $\langle W, \Sigma, \{a, b\} \rangle$. The agent's two intentions are jointly possible, since the overlapping area contains at least one world in which they both hold. However, based on the contingent facts of the situation, the agent does not believe that they will actually occur, since his beliefs, given by the set Σ , fall outside the overlap area. A rational agent will not form intentions that he does not believe can be jointly executed. Further, intentions should be nontrivial, in the sense that the agent intending ϕ should not believe that ϕ will occur without the intervening action of the agent. To enforce rationality, we define the following conditions on cognitive structures.

Definition 3.7 *A cognitive structure $\langle W, \Sigma, I \rangle$ is admissible iff it is achievable:*

$$\exists w \in \Sigma. \forall \phi \in I. w \in M_\phi$$

and nontrivial:

$$\forall \phi \in I. \exists w \in \Sigma. w \notin M_\phi.$$

This condition leads immediately to the following consequences.

⁷In terms of Chellas' minimal models, the semantics of I^* is for models that are closed under intersection. This makes sense: I^* represents any intersection of intentions.

Proposition 3.3 *These sentences are valid in all admissible structures.*

$\neg I(a \wedge \neg a)$	<i>Consistency</i>
$I(a) \wedge I(b) \supset \Diamond(a \wedge b)$	<i>Joint Consistency</i>
$I^*(a) \supset \Diamond a$	
$I^*(a) \supset B\Diamond a$	<i>Realism</i>
$I(a) \supset \neg B(\neg a)$	<i>Epistemic Consistency</i>
$I(a) \wedge I(b) \supset \neg B(\neg(a \wedge b))$	<i>Joint Epistemic Consistency</i>
$I^*(a) \supset \neg B\neg(a)$	
$I(a) \supset \neg B(a) \wedge \neg B(\neg a)$	<i>Epistemic Indeterminacy</i>

A rational agent, characterized by achievable structures, does not believe that his joint intentions represent an impossible situation: this is the theorem of Joint Epistemic Consistency. This theorem can be stated using either reading of intention.

In addition, the nontriviality condition on models means that the agent does not believe that any one of his intentions will take place without his efforts (Epistemic Indeterminacy). Recall that the B operator represents all futures the agent believes might occur, including those in which he performs various actions or does nothing. The beliefs form a background of all the possibilities among which the agent can choose by acting in particular ways. If in all these worlds a fact ϕ obtains, it does no good for an agent to form an intention to achieve ϕ , even if it is an action of the agent, because it will occur without any choice on the part of the agent. So, for example, if the agent believes he will be forced to act at some future point, perhaps involuntarily (e.g., by sneezing), it is not rational for the agent to form an intention to do that.

Note that in our logic, the realism thesis is expressed using beliefs about what is possible. This is because we distinguish beliefs about contingent facts ("Nixon was president") from the background possibilities an agent believes could occur, but haven't or won't. Realism follows directly from Joint Consistency and the simplifying assumption (1) that all worlds W are possibilities for the agent.

In this logic, we are deliberately leaving the temporal aspects vague until they are necessary. At this level of abstraction, different kinds of goals can be treated on an equal basis. For example, goals of prevention, which are problematic for some temporal logic accounts of intention, are easily represented. For an agent to prevent a state p from occurring, he must believe both p and $\neg p$ to be possible at some future state. The agent's intention is the scenario consisting of worlds in which p is always true.

3.4 Relative intentions

As we discussed earlier, one of the primary characteristics of intentions is that they are structures: agents often form intentions relative to pre-existing intentions. That is, they "elaborate" their existing plans. There are various ways in which a plan can be elaborated. For instance, a plan that includes an action that is not directly executable can be elaborated by specifying a particular way of carrying out that action; a plan that includes a set of actions can be elaborated by imposing a temporal

order on the members of the set; and a plan that includes an action involving objects whose identities are so far underspecified can be elaborated by fixing the identities of one or more of the objects. As Bratman [Bratman, 1987, p.29] notes, "[p]lans concerning ends embed plans concerning means and preliminary steps; and more general intentions ... embed more specific ones." The distinction between these two kinds of embedding recurs in the AI literature. For instance, Kautz [Kautz, 1990] identifies two relations: (1) *decomposition*, which relates a plan to another plan that constitutes a way of carrying it out (means and preliminary steps), and (2) *abstraction*, which relates a specific plan to a more general one that subsumes it. It is useful to have a term to refer to the inverse relation to abstraction: we shall speak of this as *specialization*.

Both kinds of elaboration are represented in the cognitive structure by a graph among intentions. The graph represents the means-ends structure of agent intentions. For example, suppose the agent intends to do a by doing b and c . Then the cognitive structure contains the graph fragment $M_b, M_c \rightarrow M_a$. As usual, in the cognitive structure we let the propositions stand for their associated scenarios.

Definition 3.8 *An elaborated cognitive structure consists of a cognitive structure and an embedding graph \rightarrow among intentions: $\langle W, \Sigma, I, \rightarrow \rangle$. The graph is acyclic and rooted in the primary intentions.*

Remarks. The reason we need both primary intentions and the graph structure is that, while every root of the graph must be a primary intention, primary intentions can also serve as subordinate intentions. Consider the masochistic agent with a tooth cavity: he both intends to feel pain, and intends to get his tooth filled. His cognitive structure would be:

$$\{W, \{a \supset b\}, \{a, b\}, a \rightarrow b\}.$$

Also note that a scenario of the graph may serve to elaborate more than one intention; Pollack [Pollack, 1991] calls this overloading.

The embedding graph \rightarrow is the most strongly representationalist feature of the model. It represents the structure of intentions in a direct way, by means of a relation among the relevant scenarios. A normal modal logic is incapable of this, because its accessibility relation goes from a single world (rather than a scenario) to a set of possible worlds.

In the language, \mathcal{L}_I is extended to include a modal operator $By(\alpha; \beta_1, \dots, \beta_n)$, where the β_i together are an elaboration of α .

Definition 3.9

$$\langle W, \Sigma, I, \rightarrow \rangle \models By(\alpha; \beta_1, \dots, \beta_n) \text{ iff } \beta_1, \dots, \beta_n \rightarrow \alpha.$$

For rational agents, intention elaborations will have the same properties vis-a-vis belief as top-level intentions. So, in admissible structures we insist on the condition that any scenario of \rightarrow is part of the achievable and nontrivial intentions.

Definition 3.10 *A cognitive structure $\langle W, \Sigma, I \rangle$ is admissible iff it is achievable:*

$$\exists w \in \Sigma. \forall \phi \in (I \text{ and } \rightarrow). w \in M_\phi$$

Possible Worlds W

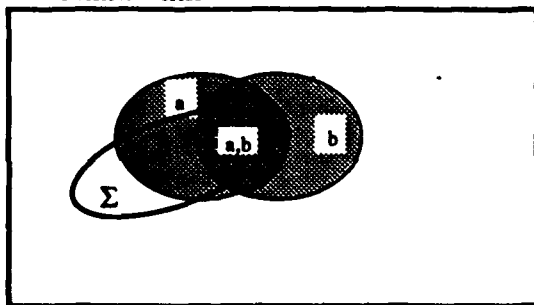


Figure 4: Means-ends intentions and belief.

and nontrivial:

$$\forall \phi \in (I \text{ and } \rightarrow). \exists w \in \Sigma. w \notin M_\phi.$$

This semantic constraint has the immediate consequence that all *By*-arguments are conjoined intentions, and share in all their properties.

Proposition 3.4

$$\models By(\alpha; \beta_i) \supset I^*(\alpha) \wedge I^*(\beta_1) \wedge \dots \wedge I^*(\beta_n)$$

But there is an additional constraint on the elaboration of intentions, having to do with their means-end relation. An agent should believe that if the elaboration is achieved, the original intention will be also. Consider the diagram of Figure 4, in which the agent has the intention to achieve *a* by achieving *b*; for concreteness, take the example of calling the telephone operator by dialing 0. There can be possible worlds in which *b* does not lead to *a*: for example, in using the internal phone system of a company. The correct rationality condition for an agent is that he believe, in the particular situation at hand, that achieving *b* will achieve *a*. This is represented by the set Σ of belief worlds, in which $b \supset a$ holds.

We call a model *embedded* if it satisfies this constraint on belief and intention structure.

Definition 3.11 A cognitive structure is embedded iff whenever $b_1 \dots b_n \rightarrow a$, $\bigcap_{i=1}^n M_{b_i} \subseteq M_a$.

It can be easily verified that this condition leads to the following theorem.

Proposition 3.5 In all embedded cognitive structures $(W, \Sigma, I, \rightarrow)$,

$$(W, \Sigma, I, \rightarrow) \models By(\alpha; \beta_1, \dots, \beta_n) \supset B(\beta_1 \wedge \dots \wedge \beta_n \supset \alpha).$$

While the embedding graph semantics is simple, it leads to interesting interactions in the statics of intention and belief. For example, in plan-recognition it can be used to determine if a recognized plan is well-formed. It is also critical to the theory of the dynamics of intention and belief. We have a preliminary theory of this dynamics expressed as a default system.

4 Conclusion

We have concentrated on the static relation between intention and belief, and shown how the relationship between these two can be represented simply by an appropriate semantics. The static formalism is useful in

task such as plan recognition, in which one agent must determine the mental state of another using partial information.

More complex applications demand a *dynamic* theory, which is really a theory of belief and intention revision. The formalism of cognitive structures can be extended readily to time-varying mental states, by adding a state index to the model. However, the theory of revision is likely to be complicated, even more so than current belief revision models [Gärdenfors and Makinson, 1990], and will probably involve elements of default reasoning.

References

- [Bratman, 1987] Michael E. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.
- [Chellas, 1980] B. F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.
- [Cohen and Levesque, 1990a] Philip R. Cohen and Hector Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990.
- [Cohen and Levesque, 1990b] Philip R. Cohen and Hector Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990.
- [Gärdenfors and Makinson, 1990] P. Gärdenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In M. Vardi, editor, *Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann, 1990.
- [Kautz, 1990] Henry A. Kautz. A circumscriptive theory of plan recognition. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, MA, 1990.
- [Konolige and Pollack, 1989] Kurt Konolige and Martha Pollack. Ascribing plans to agents: Preliminary report. *IJCAI*, Detroit, MI, 1989.
- [Levesque, 1984] Hector J. Levesque. A logic of implicit and explicit belief. *AAAI*. University of Texas at Austin, 1984.
- [Moore, 1980] Robert C. Moore. *Reasoning about Knowledge and Action*. PhD thesis, MIT, Cambridge, MA, 1980.
- [Pollack, 1991] Martha E. Pollack. Overloading intentions for efficient practical reasoning. *Nous*, 1991.
- [Rao and Georgeff, 1991] Anand S. Rao and Michael P. Georgeff. Modelling rational agents within a bdi-architecture. *KR91*, Cambridge, MA, 1991.
- [Shoham, 1990] Yoav Shoham. Agent-oriented programming. Technical Report STAN-CS-90-1335, Stanford University, Palo Alto, CA, 1990.

Ideal introspective belief

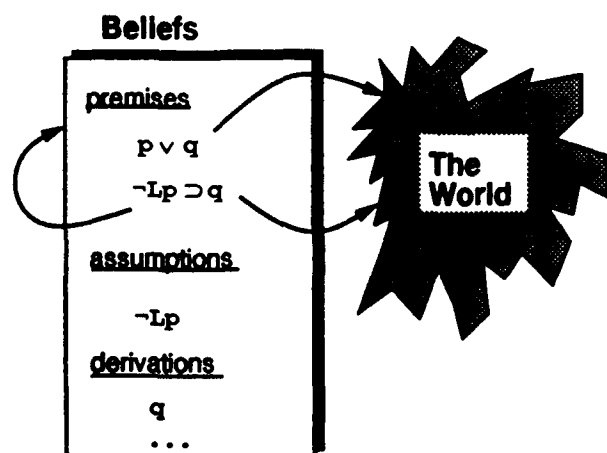
Kurt Konolige*
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
konolige@ai.sri.com

Abstract

Autoepistemic (AE) logic is a formal system characterizing agents that have complete introspective access to their own beliefs. AE logic relies on a fixed point definition that has two significant parts. The first part is a set of assumptions or hypotheses about the contents of the fixed point. The second part is a set of reflection principles that link sentences with statements about their provability. We characterize a family of ideal AE reasoners in terms of the minimal hypotheses that they can make, and the weakest and strongest reflection principles that they can have, while still maintaining the interpretation of AE logic as self-belief. These results can help in analyzing metatheoretic systems in logic programming.

Introduction

What kind of introspective capability can we expect an ideal agent to have? This question is not easily answered, since it depends on what kind of model we take for the agent's representation of his own beliefs. Autoepistemic logic (Moore [10]) uses a sentential or list semantics, which looks like this:



The beliefs of the agent are represented by sentences in a formal language. For simplicity, we consider just a propositional language \mathcal{L}_0 , and a modal extension \mathcal{L}_1 which has modal atoms of the form $L\phi$, where ϕ is a sentence of \mathcal{L}_0 .

The arrow indicates that the intended semantics of the beliefs from \mathcal{L}_0 is given by the real world, e.g., the belief q is the agent's judgment that q is true in the real world. Of course an agent's beliefs may be false, so that in fact q may not hold in the world. On the other hand, beliefs of the form $L\phi$ refer to the agent's knowledge of his own beliefs, so the semantics is just the belief set itself.

An agent starts with an initial set of beliefs, the *premises*. Through assumptions and derivations, he accumulates further beliefs, arriving finally at a belief set that is based on the premises. In order for an agent to be ideally introspective, the belief set Γ must satisfy the following equations:

$$\begin{aligned} &\text{The premises are in } \Gamma. \\ &\phi \in \Gamma \text{ and } \phi \in \mathcal{L}_0 \rightarrow L\phi \in \Gamma \\ &\phi \notin \Gamma \text{ and } \phi \in \mathcal{L}_0 \rightarrow \neg L\phi \in \Gamma \end{aligned} \quad (1)$$

Any set Γ from \mathcal{L}_1 that satisfies these conditions, and is closed under tautological consequence, will be called \mathcal{L}_1 -stable (or simply stable) for the premises Γ . The definition and term "stable set" are from Stalnaker [13]. The beliefs are stable in the sense that an agent has perfect knowledge of his own beliefs according to the intended semantics of L , and cannot infer any more atoms of the form $L\phi$ or $\neg L\phi$.

Although an ideal agent's beliefs will be a stable set containing his beliefs, not just any such set will do. For example, if the premises are $\{p \vee q\}$, one stable set is $\{p \vee q, p, Lp, L(p \vee q), \dots\}$. This set contains the belief p , which is unwarranted by the premises. The constraint of making the belief set stable guarantees that the beliefs will be introspectively complete, but it does not constrain them to be soundly based on the premises. Moore recognized this situation in formulated autoepistemic logic; his solution was to ground the belief set by making every element derivable from the premises and some assumptions about beliefs. The reason he needed a set of assumptions is that negative

*The research reported in this paper was supported by the Office of Naval Research under Contract No. N00014-89-C-0095.

introspective atoms (of the form $\neg L\phi$) are not soundly derivable from the premises alone. For example, consider the premise set $\{\neg Lp \supset q, p \vee q\}$. We would like to conclude $\neg Lp$, since there is no reasonable way of coming to believe p . But an inference rule that would allow us to conclude $\neg Lp$ would have to take into account all possible derivations, including the results of its own conclusion. This type of circular reasoning can be dealt with by adding a set of assumptions about what we expect *not* to believe, and checking at the end of all derivations that these assumptions are still valid.

In autoepistemic logic, a belief set T is called *grounded in premises* A if all of its members are tautological consequences of $A \cup LT_0 \cup \neg L\bar{T}_0$, where $LT_0 = \{L\phi \mid \phi \in T \cap \mathcal{L}_0\}$, and $\neg L\bar{T}_0 = \{\neg L\phi \mid \phi \in \mathcal{L}_0 \text{ and } \phi \notin T\}$. This concept of groundedness is fairly weak, since it relies not only on assumptions about what isn't believed ($\neg L\bar{T}_0$), but also about what is (LT_0). In this paper we consider belief sets that use only assumptions $\neg L\bar{T}_0$ in forming the belief set T . Everything else in the belief set will follow deductively (and monotonically) from the premises A and the assumptions $\neg L\bar{T}_0$. In some sense $\neg L\bar{T}_0$ is the minimal set of assumptions that we can use in this manner; for every smaller set, we have to resort to nonmonotonic rules, such as negation-as-failure [6], in order to form a stable set. For this reason we call a belief set grounded in A and $\neg L\bar{T}_0$ *ideally grounded*.

Ideally grounded logics are similar to the modal nonmonotonic logics defined in [8, 12, 7], but allow an agent to make fewer assumptions about his own beliefs. The main difference is that ideally grounded logics are more grounded in the premises than modal nonmonotonic logics, and in general will have fewer unmotivated extensions (see Section).

In the rest of this paper we explore ideally grounded belief sets from the perspective of introspective reflection principles. We are able to characterize the minimal set of principles that will yield a stable set of beliefs, and also (once nested belief operators are introduced) the maximal ones. The resultant family of introspective logics fill in a hierarchy between strongly and moderately grounded autoepistemic logic [5], and suggest that the moderately grounded fixed-point is the best system for an ideal agent with perfect awareness of his beliefs.

Minimal ideal introspection

In this and the following section we restrict the language to \mathcal{L}_1 , containing no nesting of the belief operator. This presents a simple system to explore the consequences of ideal introspection. In Section we relax this restriction and consider the fully nested modal language \mathcal{L} .

An ideally grounded introspective agent determines his belief set using the following fixed-point equation:

$$T = \{\phi \mid A \cup \neg L\bar{T}_0 \vdash_S \phi\}, \quad (2)$$

where S is some system of inference rules. Any set T that satisfies this equation will be called an *ideally*

grounded extension of A . The set $T_0 = T \cap \mathcal{L}_0$ is the *kernel* of T .

In the remainder of this section we consider the minimal set of rules S that guarantees a stable belief set for T . Because a stable set is closed under tautological consequence, the rules S must contain a complete set of propositional rules. In addition, whenever ϕ is in the belief set, we want to infer $L\phi$. The following two rules fulfill these conditions.

Rule Taut. From the finite set of sentences X infer ϕ , if ϕ is a tautological consequence of X .

Rule Reflective Up. From ϕ infer $L\phi$, if $\phi \in \mathcal{L}_0$.

Proposition 1 *Let RN be the rules Taut and Reflective Up. Every RN -extension of A is a \mathcal{L}_0 stable set containing A .*

Proof. Every extension is closed under tautological consequence by rule Taut, and the premises must be in it, by the properties of \vdash . The condition $\phi \in \Gamma$ and $\phi \in \mathcal{L}_0 \rightarrow L\phi \in \Gamma$ holds because of rule Reflective Up. The condition $\phi \notin \Gamma$ and $\phi \in \mathcal{L}_0 \rightarrow \neg L\phi \in \Gamma$ holds since any proposition ϕ not in T will be part of the assumptions $\neg L\bar{T}_0$. ■

Proposition 2 *If for every set $A \subseteq \mathcal{L}_1$, the S -extension of A is an \mathcal{L}_1 stable set containing A , then Taut and Reflective Up are admissible rules of S .*

Proof. If Taut is not an admissible rule for some extension T , then it cannot be closed under tautological consequence, and is not a stable set. Similarly, if Reflective Up is not admissible, T will contain ϕ and will not contain $L\phi$ for some proposition ϕ . ■

These two propositions show that the rules RN form the minimal logic for ideally grounded agents, in the sense that RN extensions produce stable belief sets, and they must be included in any system that produces such sets. Further, every RN extension of A is *minimal for A* : there is no stable set S containing A such that $S_0 \subset T_0$.

Proposition 3 *Every RN extension of A is a minimal stable set for A .*

Proof. Suppose there is a stable set U for A whose kernel is a proper subset of T 's. Then U must also satisfy the fixed-point condition, since the rules Reflective Up and Taut are admissible for stable sets (Proposition 2). By hypothesis the set $\neg L\bar{U}_0$ contains $\neg L\bar{T}_0$, and so U_0 must contain every element of T_0 , a contradiction. ■

The proof of this proposition points to a more general result for any class of rules that are sound with respect to the stable set conditions. An inference rule is sound with respect to stable sets if, whenever its antecedents are contained in a stable set, its consequent also must be (e.g., Reflective Up is sound because if ϕ is in a stable set, $L\phi$ must be also).

Proposition 4 *If the rules S are sound, then any S -extension of A is a minimal stable set for A .*

Proof. Suppose there is a stable set U for A whose kernel is a proper subset of T 's. Then U must also satisfy the fixed-point condition, since the rules S are admissible for stable sets. By hypothesis the set $\neg L\bar{U}_0$ contains $\neg L\bar{T}_0$, and so U_0 must contain every element of T_0 , a contradiction. ■

Groundedness, autoepistemic and default logic

In this section we relate ideally grounded extensions to their close relatives, default logic and AE extensions. Ideal groundedness is somewhat weaker than default logic and strongly grounded AE extensions, but stronger than moderately grounded ones.

Simple as it is, the system RN is almost equivalent to default logic [11]. It is not quite as strongly grounded as the latter; for while there exists a translation from DL to RN that preserves extensions, the inverse translation fails in a few cases.

We will assume that the reader is familiar with DL. A default theory $\langle W, D \rangle$ consists of a set of first-order sentences W and a set of defaults D of the form

$$\alpha : \beta_1, \dots, \beta_n / \gamma.$$

Here only the propositional case will be considered, but extending the results to first-order languages is straightforward (as long as no quantifying-in is allowed, e.g., sentences of the form $Qx.L\phi(x)$).

To get a translation to RN, simply take W and add a translation of each default rule, as follows:

$$A = W \cup \{L(\alpha \wedge \alpha) \wedge \neg L\neg\beta_1 \dots \supset \gamma \mid \alpha : \beta_1, \dots, \beta_n / \gamma \in D\}. \quad (3)$$

Note the form of the first modal atom: $L(\alpha \wedge \alpha)$, rather than $L\alpha$. Since the beliefs of an agent are closed under tautological consequence, this amounts to the same constraint on beliefs; however, the difference is important for finding extensions, as will be made clear shortly.

Proposition 5 U is the kernel of an RN extension of A iff it is a DL extension of $\langle W, D \rangle$.

Proof. Let $A = W \cup \{L(\alpha \wedge \alpha) \wedge \neg L\neg\beta_1 \supset \gamma \mid \alpha : \beta_1, \dots, \beta_n / \gamma \in D\}$. We will show that the set

$$\Gamma(U) = \{\phi \in \mathcal{L}_0 \mid A \cup \neg L\bar{U} \vdash_{RN} \phi\}$$

is the least set satisfying the properties:

$$W \subseteq \Gamma(U).$$

2 $\Gamma(U)$ is closed under tautological consequence.

3 For $\alpha : \beta_1, \dots, \beta_n / \gamma \in D$, if $\alpha \in \Gamma(U)$ and $\neg\beta \notin U$, then $\gamma \in \Gamma(U)$.

The first two properties follow directly from the definition of $\Gamma(U)$. The third property follows by simple propositional inference, given the form of A .

To show $\Gamma(U)$ is minimal, note that it is the set of tautological consequences of W and some set γ_i of conclusions of defaults. To make it smaller, we would have to eliminate some of the γ_i . But it is clear from

the discussion below that the only way a γ_i could be present is if the third condition defining $\Gamma(U)$ holds; thus all γ_i must be present, and $\Gamma(U)$ is minimal.

We can reduce the definition of extensions (2) to use only the kernel:

$$U = \{\phi \in \mathcal{L}_0 \mid A \cup \neg L\bar{U} \vdash_S \phi\}.$$

This gives a fixed-point condition defining extensions as

$$U = \Gamma(U)$$

which is the same as for default logic. ■

This is a simple translation of DL into a minimal AE logic. It is the same as the translation in [5] (except for the use of $\alpha \wedge \alpha$ instead of α), but there it was necessary to limit the extensions of the AE logic to strongly grounded ones, a syntactic method based on the form of the premises. No such method is needed here.

The stipulation on the form of $L(\alpha \wedge \alpha)$ is necessary to prevent derivations that arise from the interaction of modal atoms. Consider the two theories:

$$\begin{aligned} &\{\neg Lp \supset p, Lp \supset p\} \\ &\{\neg Lp \supset p, L(p \wedge p) \supset p\} \end{aligned}$$

The first one has an RN extension $Cn(p)$, because p is a tautological consequence of the initial constraints. On the other hand, it is not a consequence of the second set of constraints, because $\neg Lp$ and $L(p \wedge p)$ are consistent from the view of propositional logic. Since there is no way to derive p by any of the rules, $Cn(p)$ cannot be an extension; yet assuming $\neg Lp$ leads to the derivation of p and a contradiction. So the second set has no extensions.

To get autoepistemic logic, we need to include more assumptions about beliefs in the fixed point equation 2. Let us define *open RN extensions* as solutions of the equation

$$T = \{\phi \mid A \cup LT_0 \cup \neg L\bar{T}_0 \vdash_{RN} \phi\}, \quad (4)$$

where LT_0 is the set $\{L\phi \mid \phi \in T_0\}$. Actually, the presence of the Up rule is redundant here. From results in [5], it is easy to show the following proposition.

Proposition 6 T is an open RN extension of A iff it is the kernel of an AE extension of A .

The kernel of an AE extension is just the part of the extension from \mathcal{L}_0 . The kernel completely determines the extension.

So the basic difference between AE and default logic is based on the groundedness of the extensions, that is, AE logic lets an agent assume belief in a proposition α , and use that assumption to derive the very same proposition as part of the final set of beliefs. In default logic, all derivations must be ideally grounded, so that assumptions are of the form $\neg L\phi$.

The circular reasoning possible in AE logic was noted in [5], and two increasingly stronger notions, moderate and strong groundedness, were defined as a means

of throwing out extensions that exhibit such reasoning. Moderately grounded extensions of A are defined as those AE extensions are also minimal stable sets containing A . Strongly grounded extensions use a syntactic method to eliminate all inferences from facts to belief propositions, e.g., even with the premise set

$$A = \{La \supset a, \neg La \supset a\} \quad (5)$$

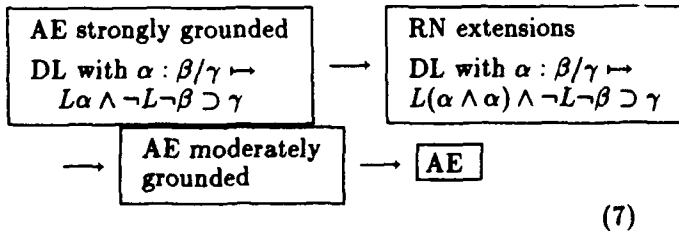
there is no derivation of a , because La and $\neg La$ are not allowed to interact. This means that different sets A , even if they are propositionally equivalent, can generate different extensions. Strongly grounded extensions are equivalent to default logic extensions under the simple translation of default rules:

$$\alpha : \beta_1, \dots, \beta_n / \gamma \mapsto L\alpha \wedge \neg L\neg\beta_1 \wedge \dots \wedge \neg L\neg\beta_n \supset \gamma. \quad (6)$$

Note the difference with the translation of (3): $L\alpha$ instead of $L(\alpha \wedge \alpha)$.

Here, rather than defining restrictions on extensions, we have taken the approach of trying to find the minimal reflective principles that will allow an agent full knowledge of his beliefs, at the same time trying to make them as grounded as possible. The result is a logic that is somewhere between moderately and strongly grounded AE extensions, and which can imitate the groundedness conditions of default logic.

Let us define one fixed point logic $S1$ to be included in another $S2$ ($S1 \rightarrow S2$) if for any premise set the extensions of $S1$ are always extensions of $S2$, and for some premise set there is an extension of $S2$ that is not an extension of $S1$. $S1$ is the stronger nonmonotonic logic if we define ϕ as a consequence of a premise set just in case ϕ is in every extension of the premises. The relationship among the various AE logics can be diagrammed as follows:



(7)

Nested belief

So far we have preferred to forego the complications of beliefs about beliefs, using the language \mathcal{L}_1 that contains no nesting of modal operators. This language and its semantics can be extended in a straightforward way. Let \mathcal{L} be the propositional modal language formed from \mathcal{L}_0 by the recursive addition of atoms of the form $L\mu$, with $\mu \in \mathcal{L}$.

The semantic equations for a stable set (1) are modified to take away the restriction of beliefs being in \mathcal{L}_0 :

$$\begin{aligned} &\text{The premises are in } \Gamma. \\ &\phi \in \Gamma \rightarrow L\phi \in \Gamma \\ &\phi \notin \Gamma \rightarrow \neg L\phi \in \Gamma \end{aligned} \quad (8)$$

Any set from \mathcal{L} that satisfies these conditions, and is closed under tautological consequence, will be called a stable set for A (in contrast to \mathcal{L}_1 -stable, which does not consider nested modal atoms).

Consider a premise set A that is drawn from \mathcal{L}_1 , as before. In every RN extension of A there is complete knowledge of what facts are believed or disbelieved, i.e., $L\phi$ or $\neg L\phi$ is present for every nonmodal ϕ . The addition of the nested modal atoms should make no difference to this picture, except to reflect the presence of the belief atoms in the correct way. So, for example, if La is in an RN extension S , then LLa should be in the extension when we consider \mathcal{L} ; and similarly $L\neg La$ should be present if $\neg La$ is not in S . This much is easily accomplished by removing the restriction on Reflective Up, and giving it its usual name from modal logic.

Rule Necessitation. From ϕ infer $L\phi$.

This rule will add positive modal atoms; but we need also to add negative ones. For example, if La is in an extension, and the extension is consistent, then $\neg La$ is not in it, and this fact should be reflected in the presence of $\neg L\neg La$. In fact we want to infer $\neg L\mu$ for every sentence μ that will not be in the extension, given that we have full knowledge of the belief atoms from \mathcal{L}_1 . Suppose that there is a sentence $La \vee \neg Lb \vee c$ that is not in S , where c is a nonmodal sentence. This implies that, for stable S , $\neg La \in S$, $Lb \in S$, and $\neg Lc \in S$. So from these latter sentences we should infer $\neg L(La \vee \neg Lb \vee c)$. This is what the following rule does.

Rule Fill. From $La_i, \neg L\beta_j, \neg L\gamma$, and $\mu \supset (\bigvee_i La_i \vee \bigvee_j \neg L\beta_j \vee \gamma)$, infer $\neg L\mu$.

The system NRN consists of the rules Taut, Necessitation, and Fill. The basic properties of NRN extensions are that they are minimal stable sets, the rules are essential, and they are conservative extensions of RN fixed points.

Proposition 7 If for every set $A \subseteq \mathcal{L}$, the S -extension of A is a stable set containing A , then Taut, Necessitation, and Fill are admissible rules of S .

Proof. Taut and Nec are the same as for Proposition 2.

For Fill, note that every consistent stable set containing the premises to the rule cannot contain μ , and so must contain $\neg L\mu$. ■

Proposition 8 Every NRN extension of A is a stable set for A .

Proof. Assume that T is a consistent NRN extension of A . By rule Nested Reflective Up, the first part of the semantic definition is satisfied. For negative modal atoms, we proceed by induction on the level of nesting of L . By definition and the rule Nested Reflective Up, either $L\phi$ or $\neg L\phi$ is in T for every nonmodal ϕ . Suppose a sentence $s = (\bigvee_i La_i \vee \bigvee_j \neg L\beta_j \vee \gamma) \in \mathcal{L}_1$ is not in T . Then each of $\neg La_i, L\beta_j$ and $\neg L\gamma$ is in T . By rule Fill, $\neg L\mu$ is in T for any $\mu \supset s$. Hence for every sentence $\nu \in \mathcal{L}_1$, the negative semantic rule is

satisfied, and either $L\nu$ or $\neg L\nu$ is in T . By induction, it can be shown that the semantic rule is satisfied for all levels of nesting. ■

Extensions that are stable sets are also minimal, as for the nonnested language.

Proposition 9 *If the rules S are sound with respect to stable sets, and the S -extension of A is a stable set, then it is a minimal stable set for A .*

Proof. Same as for Proposition 4. ■

Proposition 10 *If $A \subseteq \mathcal{L}_1$, then the kernel of every RN extension is the kernel of an NRN extension, and conversely, the kernel of every NRN extension is the kernel of an RN extension.*

Proof. The converse is obvious, since the rules NRN include RN. For the original direction, assume we have an RN extension S , which contains $L\phi$ or $\neg L\phi$ for every $\phi \in \mathcal{L}_0$. From the proof of Proposition 8, it is clear that the set $T = \{\mu \mid S \vdash_{\text{NRN}} \mu\}$ is a stable set for A , and further it is an NRN extension, since all elements of its kernel are derivable from A and $\neg L\bar{S}$. ■

Finally, we can show that the Fill rule is redundant if the schema K ($[L\phi \wedge L(\phi \supset \psi)] \supset L\psi$) is present.

Proposition 11 *The rule Fill is admissible in any system containing K , Taut and Necessitation.*

Proof. Suppose each of $\neg L\alpha_i$, $L\beta_j$ and $\neg L\gamma$ is in A , together with K and all instances of K . Let $\mu = \bigwedge_i \neg L\alpha_i \wedge \bigwedge_j L\beta_j$. By Taut and Up, $L[\mu \wedge (\mu \supset \gamma)]$ is derivable, and from schema K and $\neg L\gamma$ we have $\neg L[\mu \wedge (\mu \supset \gamma)]$. Since we also have $L\mu$ by Up, this gives (using K) $\neg L(\mu \supset \gamma)$. Again by K and Taut, we could derive $\neg L\nu$ for any ν such that $\nu \supset (\mu \supset \gamma)$ is a tautology. ■

Because nested modal atoms are propositionally distinct from nonnested ones, it is possible to derive new translations from default logic to sentences of \mathcal{L} such that all extensions are strongly grounded and hence equivalent to default logic extensions. There are many ways to do this; all that is required is to translate from $\alpha : \beta/\gamma$ to a sentence in which α and β are put under different nestings of modal operators that correspond to the single nesting semantics. For example, three such translations are:

- a) $LL\alpha \wedge \neg L\neg\beta \supset \gamma$
- b) $L\alpha \wedge \neg LL\neg\beta \supset \gamma$
- c) $L\alpha \wedge L\neg L\neg\beta \supset \gamma$

Reflective reasoning principles

The systems RN and NRN are minimal rules that might be used by an agent reasoning about its own beliefs. They have the nice characteristic of giving minimal stable sets, and so are somewhere between strongly and moderately grounded. But are there other reflective reasoning principles that could be incorporated? In this

section we will give a partial answer to this question by examining several standard modal axiomatic schemata, and showing how some of them are appropriate as general reasoning principles, while others must be regarded as specific assumptions about the relation of beliefs to the world.

The most well-known modal schemata are the following.

- K. $L(\phi \supset \psi) \supset (L\phi \supset L\psi)$
- T. $L\phi \supset \phi$
- D. $L\phi \supset \neg L\neg\phi$
- 4. $L\phi \supset LL\phi$
- 5. $\neg L\phi \supset L\neg L\phi$

The first question we could ask is: which of these schemata are sound with respect to the semantics of amalgamated belief sets? It should be clear that K, 4 and 5 are all sound, since if their antecedents are true of a stable set, then so are their consequents. The schema D is true only of consistent stable sets, as we might expect, since it says that a sentence can be in a belief set only if its negation is not.

The schema T, on the other hand, is not semantically valid. It is possible for an agent to believe a fact ϕ , but that fact may not be true in the real world. Asserting T for a particular fact ϕ says something about the agent's knowledge of how his beliefs are related to the world, and causes different reasoning patterns to appear in an agent's inferences about his own beliefs.

Here is a short example of how the sentence $Lp \supset p$ could be used by an agent. Consider the propositions:

- p = The copier repairman has arrived
- q = The copier is ok

Suppose an agent believes that if he has no knowledge that the repairman has arrived, the copier must be ok. Further he believes that the copier is broken. We represent this as:

$$A = \{\neg q, \neg Lp \supset q\} . \quad (11)$$

The premises A do not have any NRN or AE extension, because while Lp is derivable, p is not. One solution is to give the agent confidence in his own beliefs, e.g.,

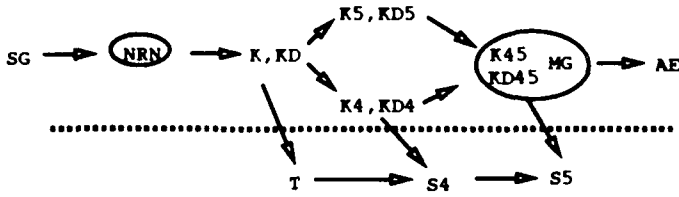
$$A' = \{\neg q, \neg Lp \supset q, Lp \supset p\} . \quad (12)$$

Now there is an NRN-extension in which p is true, since from Lp the agent can derive p . It is as if the agent says, "I believe that p , therefore p must be the case."

Although one might not want to use this type of reasoning in a particular agent design, the point is that T sanctions a certain type of reasoning about the connection of beliefs to the world, and is thus a "nonlogical" axiom, similar to $\neg Lp \supset q$.

Different modal systems can be constructed by combining the different modal schemata with the inference rules Taut and Necessitation. Using our previous definition of inclusion, we show the following relations among the different versions of S -extensions.

Proposition 12 *The following diagram gives all the inclusion relations of ideally grounded extensions based on the modal systems formed from the schemas K , T , D , 4 , and 5 .*



Proof. We will sketch the technique for two examples. The basic idea is to consider a theory containing variations of the pair of sentences $Lp \supset p$, $\neg Lp \supset p$. This theory has the single extension with kernel $Cn(p)$. For the system K , consider the pair $Lp \supset p$, $\neg L(p \wedge p) \supset p$. This theory has no RN extensions. But it does have a K -extension, since in the system K one infers p . Hence K extensions and RN extensions are distinct. For the schema 4 , consider the pair of sentences $LLp \supset p$, $\neg L(p \wedge p) \supset p$. No K or RN extensions exist; but there is a $K4$ extension, since in $K4$ the pair infers p . Similar pairs can be found for the other systems. ■

The top half are systems whose extensions are all subsets of AE logic. SG stands for strongly grounded AE extensions, and MG for moderately grounded. The minimal ideally grounded system is NRN, and the maximum is $K45$ or $KD45$, which is equivalent to MG (see [5]). An ideal introspective agent would use $KD45$ extensions, which we call ideal extensions. Note that the schema D does not make any difference as far as ideally grounded extensions are concerned; in effect, the agent cannot use reasoning about self-belief to detect an incoherence in his beliefs.

In fact all of the systems from NRN to $KD45$ are very similar. Their only difference comes from premise sets that contain sentences of the form

$$\begin{aligned} &\neg Lp \supset p \\ &\alpha \supset p, \end{aligned}$$

where $\alpha \supset Lp$ is a theorem of the modal system. For example, in K we have $L(p \wedge p) \supset Lp$, and a premise set as above with $\alpha = L(p \wedge p)$ would distinguish K from NRN, in that the former would have an extension containing p . Similarly, $\alpha = \neg L\neg Lp$ could be used for $K5$. But the sentence $\neg Lp \supset p$ is generally not one that captures a useful introspective reasoning pattern, and would probably not occur by design in an application. There thus seems to be no practical difference between NRN and $KD45$, since the additional axioms do not result in potentially interesting reasoning patterns.

The second tier is present for formal completeness. The axiom schema T , we have argued, is a useful way of characterizing a domain-dependent and proposition-dependent connection between the agent's beliefs and

the world. These systems do not respect sound autoepistemic reasoning, and are not included in AE logic: the extensions generated using instances of T can differ significantly from AE extensions. In fact, if the AE fixed-point equation (4) is supplied with the axiom schema T , then it degenerates into monotonic $S5$ [9, 10]. This is because it interacts with the positive assumptions LT_0 , producing arbitrary ungrounded beliefs. In ideally grounded logic, the T schema can serve a useful representational purpose, and all modal systems, including $S5$, produce nonmonotonic fixed points.

Modal nonmonotonic logics

Modal nonmonotonic logics are based on the following fixed point equation:

$$T = \{\phi \mid A \cup \neg L\bar{T} \vdash_S \phi\},$$

where S is a modal system. McDermott [8] analyzed this equation for the systems T , $S4$, and $S5$. Subsequent investigations [12, 7] considered many other modal systems, including most of those mentioned in this paper. The difference with ideally grounded extensions is the presence of assumptions containing nested atoms, e.g., $\neg L\neg Lp$. For an ideal agent, this amounts to an assumption of Lp , since any stable set not containing $\neg Lp$ must contain Lp . In fact, modal nonmonotonic logics whose underlying modal system contains the schema 5 are all equivalent to AE logic. And as with AE logic, the schemas 5 and T combine to collapse the fixed point to monotonic $S5$.

From the point of view of ideally grounded extensions, the assumption set $\neg L\bar{T}$ is too "large." The schema 5 , which in ideally grounded extensions is just a principle of reasoning about derived beliefs, in modal nonmonotonic logic also interacts with nested negative assumptions to produce positive ones. The inclusion diagram for ideally grounded extensions is almost the same as that for the normal modal systems serving as a deductive base (see [2]), except for the schema D . But all modal nonmonotonic logics containing the K and 5 schemas (but not T) are equivalent to weakly grounded AE logic because of their large assumption set, collapsing systems that are distinct in the ideally grounded case. Because of this, modal nonmonotonic logic misses the moderately grounded endpoint. In fact, no modal nonmonotonic logic produces only minimal stable sets: in the simplest system N , containing only the necessitation rule and no logical axioms, the premises $\{Lp \supset p, \neg L\neg Lp \supset p\}$ have two extensions, $Cn()$ and $Cn(p)$. Only the first of these is minimal.

Conclusion

We have presented the minimal logic (NRN) that an ideal introspective agent should use. It is minimal in the sense that the agent makes a minimal set of assumptions about his own beliefs, and employs a minimal set of rules necessary to guarantee that his beliefs are stable. An ideal introspective reasoner may enjoy more

powerful rules of introspection, for example the modal schemas 4 and 5, but he should keep the assumptions about his beliefs to a minimum. The schema T is not a sound axiom for an introspective agent, but can be used to characterize a contingent connection between beliefs and the world.

The concept of ideally grounded extensions first appeared in [5], where the system KD45 was presented and proven equivalent to moderately grounded AE extensions.¹ Fixpoints of the systems T, S4 and S5 were introduced under the name of nonmonotonic ground logics in [14], and it was shown that the S5 logic was nondegenerate and consistent, i.e., does not reduce to monotonic S5, and always has an extension.

Ideally grounded logic might be employed in an analysis of metatheoretic systems, such as the DEMO and SOLVE predicates in logic programming [1, 3]. Using a predicate to represent provability can cause problems with syntax and consistency (see [4] for some comments). Instead, this research suggests using a modal operator, and defining a theory by the fixed point definition (2). Some appropriate notion of negation-as-failure would be used to generate the assumptions, and the rest of the fixed point could be calculated using the reflection rules.

References

- [1] K. A. Bowen and R. A. Kowalski, Amalgamating language and metalanguage in logic programming, Computer and Information Science Report 4/81, Syracuse University (1981).
- [2] B. F. Chellas, *Modal Logic: An Introduction* (Cambridge University Press, 1980).
- [3] S. Costantini, Semantics of a metalogic programming language, *International Journal of Foundations of Computer Science* 1 (3) (1990).
- [4] J. des Rivières and H. Levesque, The consistency of syntactical treatments of knowledge, in: J. Y. Halpern, ed., *Conference on Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, 1986) 115-130.
- [5] K. Konolige, On the relation between default and autoepistemic logic, *Artificial Intelligence* 35 (3) (1988) 343-382.
- [6] J. W. Lloyd, *Foundations of Logic Programming* (Springer-Verlag, Berlin, 1987).
- [7] W. Marek, G. F. Schwarz, and M. Truszczyński, Modal nonmonotonic logics: ranges, characterization, computation, in: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, Cambridge, MA (1991).
- [8] D. McDermott, Non-monotonic logic II, *Journal of the ACM* 29 (1982) 33-57.
- [9] D. McDermott and J. Doyle, Non-monotonic logic I, *Artificial Intelligence* 13 (1-2) (1980) 41-72.
- [10] R. C. Moore, Semantical considerations on non-monotonic logic, *Artificial Intelligence* 25 (1) (1985).
- [11] R. Reiter, A logic for default reasoning, *Artificial Intelligence* 13 (1-2) (1980).
- [12] G. F. Schwarz, Autoepistemic modal logics, in: *Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA (1990).
- [13] R. C. Stalnaker, A note on nonmonotonic modal logic, Department of Philosophy, Cornell University, (1980).
- [14] M. Tiomkin and M. Kaminski, Nonmonotonic default modal logics, in: *Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA (1990).

¹A slightly different fixed-point was used because of a technical difference in the form of monotonic inference in modal systems.